Day 5: Data visualization and communication

Principles of Data Visualization

Sebastian Ramirez-Ruiz
Hertie School

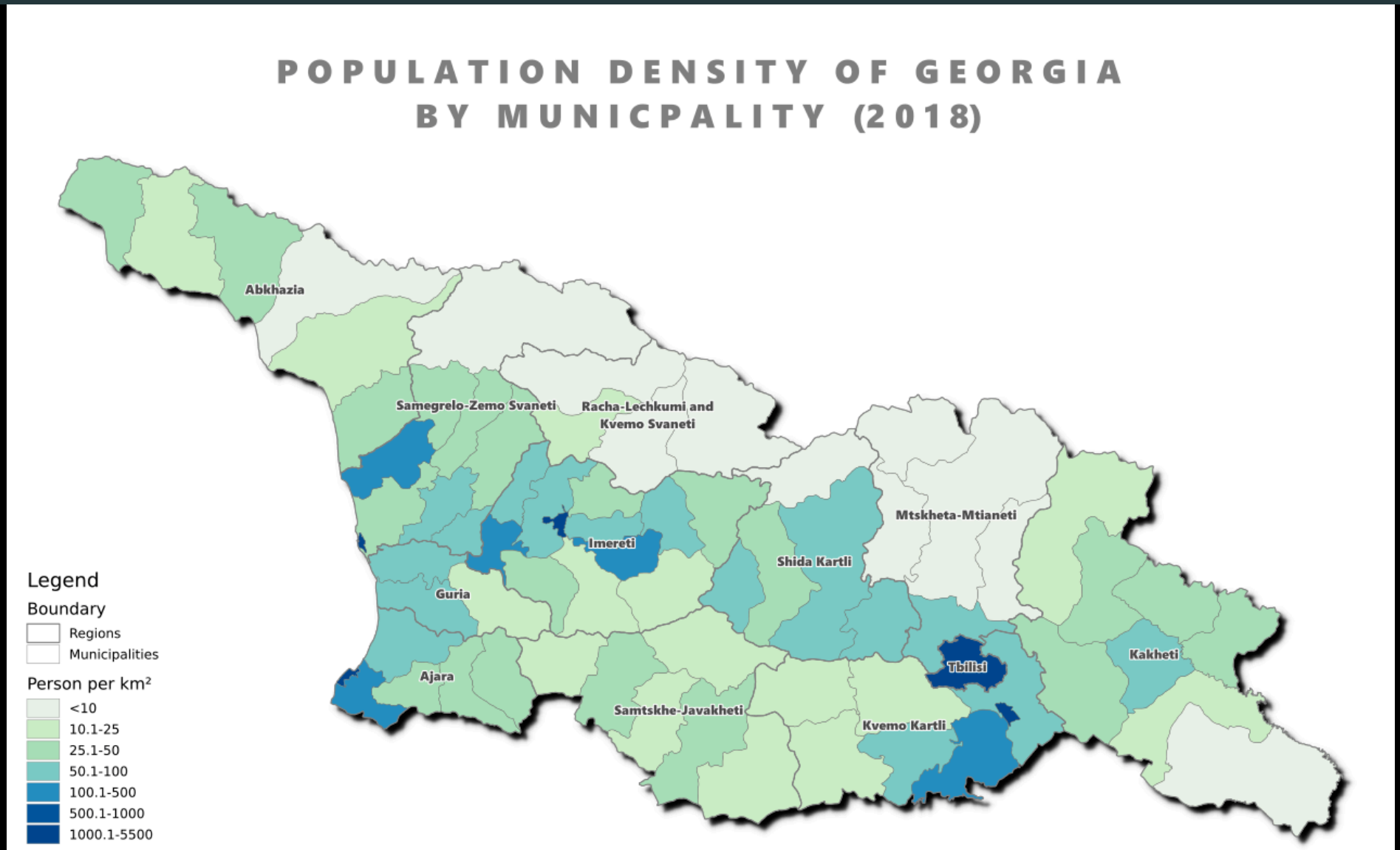
Table of contents

[1] This section draws on materials from Claus Wilke's excellent book *Fundamentals of Data Visualization*.

Any ideas?

490801	42.77652777795696,42.2804166666433,1.169209
490802	42.77680555573473,42.2804166666433,1.169209
490803	42.79041666684586,42.2804166666433,1.169209
490804	42.79097222240141,42.2804166666433,1.169209
490805	42.79125000017919,42.2804166666433,1.169209
490806	42.791527777956965,42.2804166666433,1.169209
490807	42.79180555573475,42.2804166666433,1.169209
490808	42.870138889068144,42.2804166666433,1.169209
490809	42.98291666684601,42.2804166666433,1.169209
490810	42.983472222401566,42.2804166666433,1.169209
490811	42.98458333351268,42.2804166666433,1.169209
490812	43.07847222240164,42.2804166666433,1.169209
490813	43.07986111129053,42.2804166666433,1.169209
490814	43.08041666684609,42.2804166666433,1.169209
490815	43.08819444462387,42.2804166666433,1.169209
490816	43.09680555573499,42.2804166666433,1.169209
490817	43.09708333351277,42.2804166666433,1.169209
490818	43.09902777795722,42.2804166666433,1.169209
490819	43.09930555573499,42.2804166666433,1.169209
490820	43.10375000017944,42.2804166666433,1.348093
490821	43.104305555734996,42.2804166666433,1.348093

How about now?

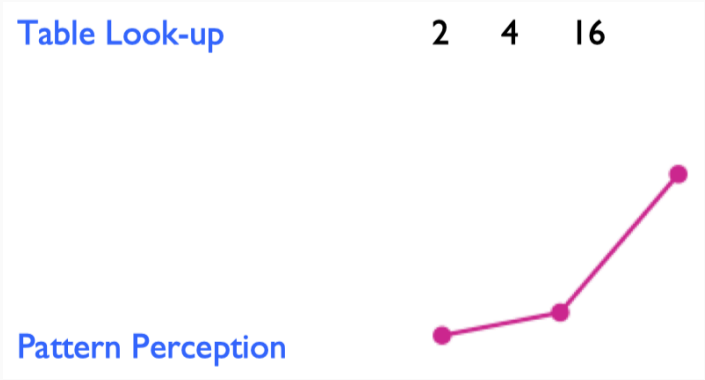


Why data visualization?

Why data visualization?

A new method for the policy toolbox

- Data visualization is a method for **making sense** (*and not just pictures*) of data.
- Note that this is more than data visualization in the narrow sense, i.e. the act of *encoding quantitative information* in visual objects.
- Decision-makers are mostly interested in *patterns*, not individual and exact values.
- Two ways to make sense of quantitative information:

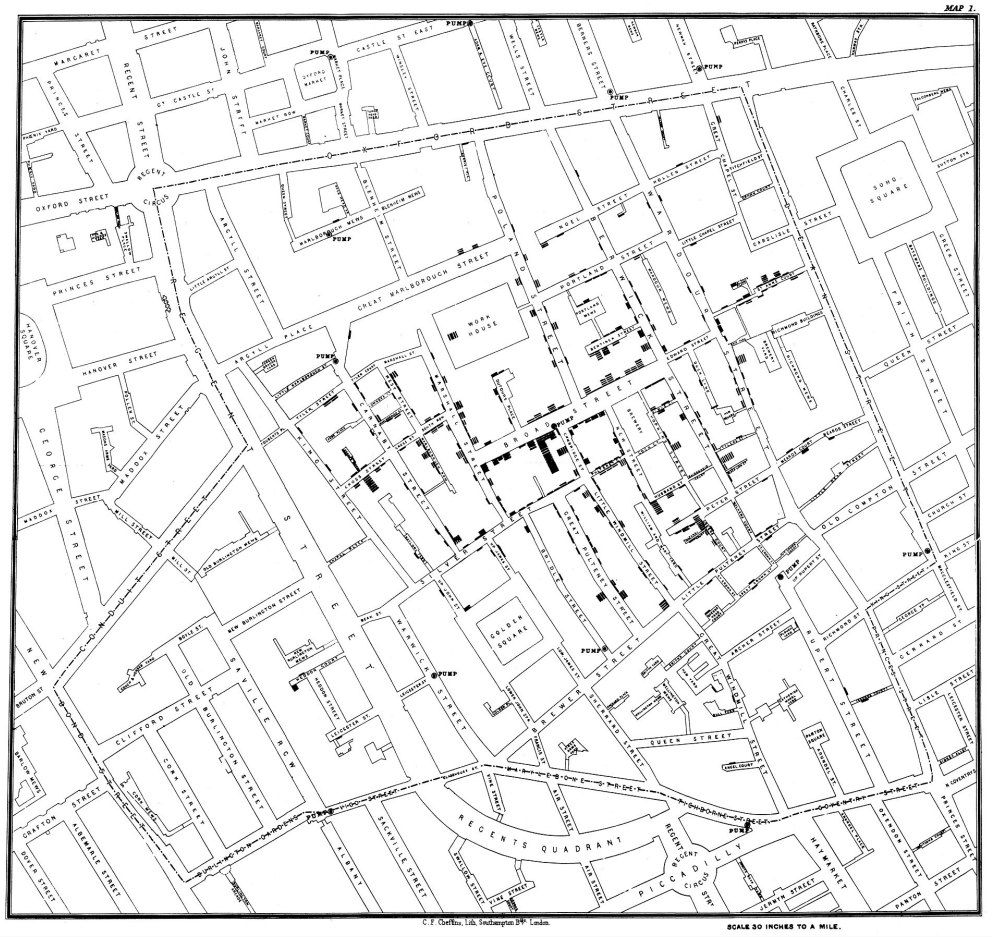


The case for visualization

- Visualization **provides useful summaries** for large, complicated data sets – in fact, the utility of visualization increases with data size.
- Visualization **lets you see things** that would otherwise be invisible, in particular relationships among data (patterns, trends, exceptions).
- Visualization comes with **little or no assumptions** about the nature of the data.
- Visualization facilitates interaction between researcher and data – **it's a hypothesis generating device**.

"The critical question is how best to transform the data into something that people can understand for optimal decision making." Colin Ware, 2013

1854 Broad Street cholera outbreak



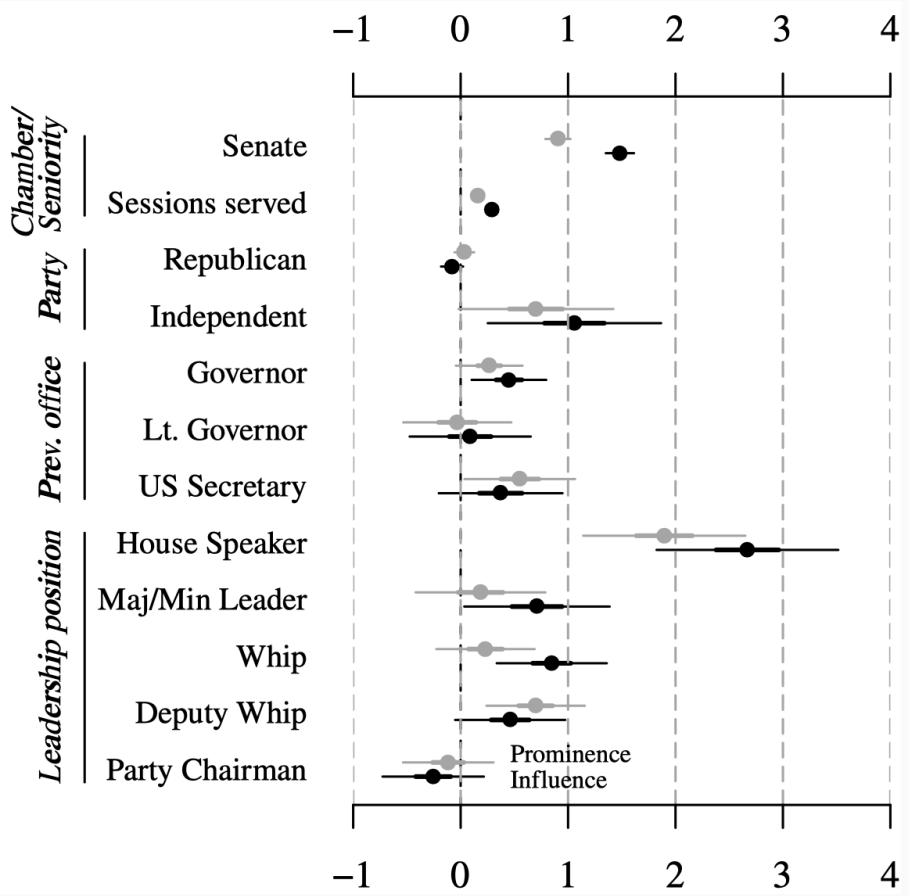
- One of the most famous data viusalizations of all times: John Snow's cholera case map.
- The Broad Street cholera outbreak in in Soho, London in 1854 was studied by physician John Snow to study its causes (rival hypotheses: germ-contaminated water vs. airborne transmission).
- The germ theory was not established at this point but the map helped highlight how cases clustered around a contaminated pump (which was by far not the only source of contaminated water though).
- Fun fact: this is what doing good data viz gives you:

Graphs vs. tables I: model estimates

Hertie School

	Prominence	Influence
Senate	0.906***	1.483***
	(0.060)	(0.067)
Sessions served	0.163***	0.292***
	(0.016)	(0.017)
Party (Independent)	0.701*	1.059**
	(0.368)	(0.412)
Party (Republican)	0.035	−0.080
	(0.047)	(0.052)
Office: Governor	0.266*	0.450**
	(0.158)	(0.177)
Office: Lt. Governor	−0.031	0.089
	(0.257)	(0.288)
Office: US Secretary	0.551**	0.372
	(0.262)	(0.294)
Position: House Speaker	1.896***	2.670***
	(0.385)	(0.431)
Position: Majority/Minority Leader	0.185	0.711**
	(0.308)	(0.345)
Position: Whip	0.231	0.848***
	(0.233)	(0.261)
Position: Deputy Whip	0.698***	0.462*
	(0.234)	(0.262)
Position: Party Chairman	−0.115	−0.255
	(0.215)	(0.241)
(Intercept)	1.648***	1.527***
	(0.050)	(0.057)
N	492	492
R-squared	0.493	0.694
Adj. R-squared	0.481	0.687
Residual Std. Error (df = 479)	0.505	0.565
F Statistic (df = 12; 479)	38.890***	90.715***

***p < .01; **p < .05; *p < .1



Graphs vs. tables II: amounts

Country	Length of Constitution
Bosnia and Herzegovina	5,230
Montenegro	7,074
Andorra	8,740
Macedonia	9,231
Croatia	10,898
Slovenia	11,410
Italy	11,708
Albania	13,747
Spain	17,608
Serbia	19,891
Greece	27,177
Malta	31,820
Portugal	35,181

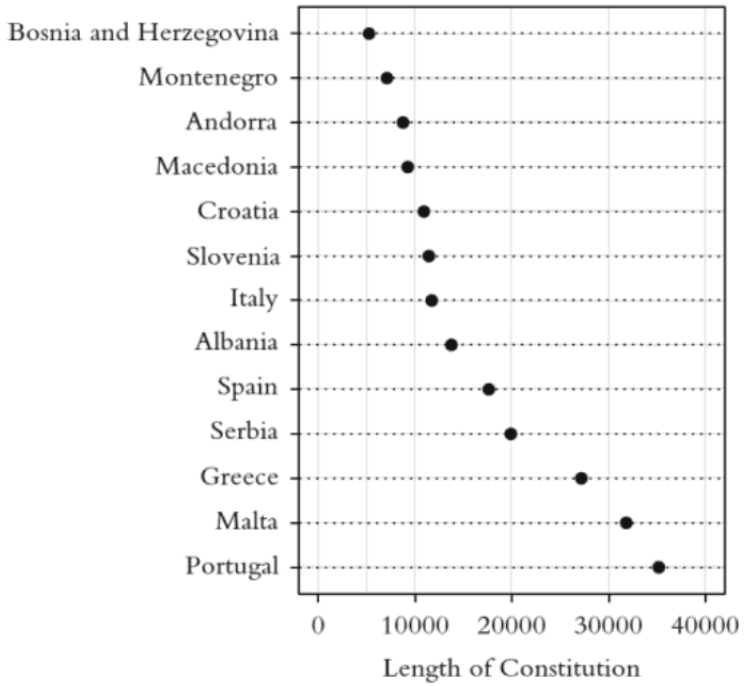


Figure 10.2 Length (in words) of present-day constitutions in countries in Southern Europe. Although it is possible from the table to observe the patterns that jump out in the graph—for example, the large difference between the shortest and longest constitutions—it requires far more (unnecessary) cognitive work.[50]

Graphs vs. tables III: relationships

- The table on the right comprises data sets I through IV, each consisting of eleven (x, y) points.
- **Carefully study the table**. How do x and y as well as their relationship compare across datasets?

I.		II.		III.		IV.	
y1	*x1*	*y2*	*x2*	*y3*	*x3*	*y4*	*x4*
8.04	10	9.14	10	7.46	10	6.58	8
6.95	8	8.14	8	6.77	8	5.76	8
7.58	13	8.74	13	12.74	13	7.71	8
8.81	9	8.77	9	7.11	9	8.84	8
8.33	11	9.26	11	7.81	11	8.47	8
9.96	14	8.10	14	8.84	14	7.04	8
7.24	6	6.13	6	6.08	6	5.25	8
4.26	4	3.10	4	5.39	4	12.50	19
10.84	12	9.13	12	8.15	12	5.56	8
4.82	7	7.26	7	6.42	7	7.91	8
5.68	5	4.74	5	5.73	5	6.89	8

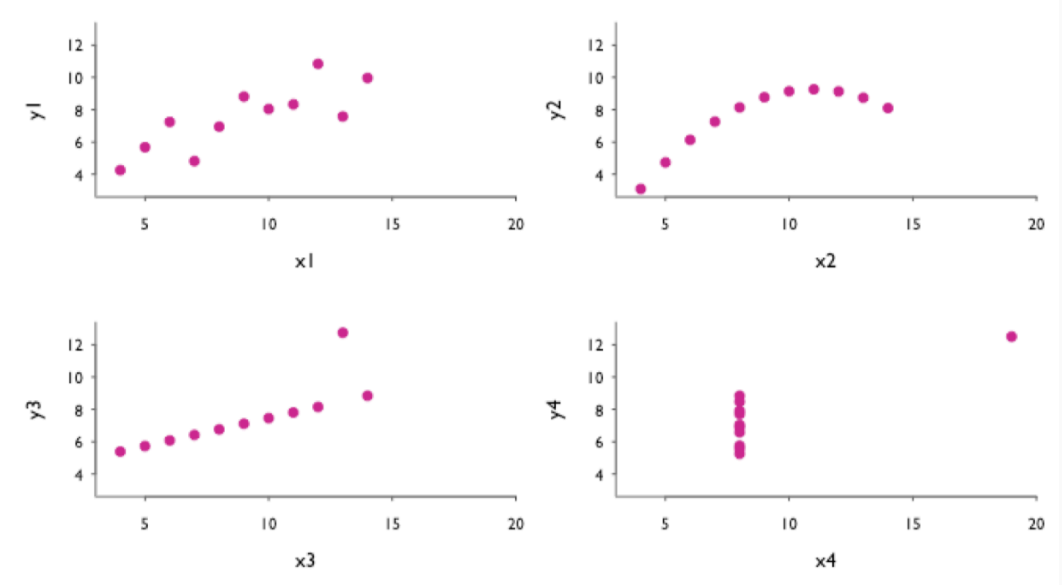
Graphs vs. tables III: relationships (cont.)

- The table on the right comprises data sets I through IV, each consisting of eleven (x, y) points.
- **Carefully study the table**. How do x and y as well as their relationship compare across datasets?
- It shows that all the data sets have nearly identical simple descriptive statistics in terms of mean, standard deviation, correlation, and linear fit!

	I.		II.		III.		IV.	
	y1	*x1*	*y2*	*x2*	*y3*	*x3*	*y4*	*x4*
	8.04	10	9.14	10	7.46	10	6.58	8
	6.95	8	8.14	8	6.77	8	5.76	8
	7.58	13	8.74	13	12.74	13	7.71	8
	8.81	9	8.77	9	7.11	9	8.84	8
	8.33	11	9.26	11	7.81	11	8.47	8
	9.96	14	8.10	14	8.84	14	7.04	8
	7.24	6	6.13	6	6.08	6	5.25	8
	4.26	4	3.10	4	5.39	4	12.50	19
	10.84	12	9.13	12	8.15	12	5.56	8
	4.82	7	7.26	7	6.42	7	7.91	8
	5.68	5	4.74	5	5.73	5	6.89	8
Mean(*y*)		7.50		7.5		7.50		7.5
Mean(*x*)		9.0		9.0		9.0		9.0
SD(*y*)		2.03		2.03		2.03		2.03
SD(*x*)		3.32		3.32		3.32		3.32
r(*y*, *x*)		.82		.82		.82		.82
y = a + bx		y = 3 + 0.5*x*		y = 3 + 0.5*x*		y = 3 + 0.5*x*		y = 3 + 0.5*x*
R^2		.67		.67		.67		.67

Graphs vs. tables III: relationships (cont.)

- The table on the right comprises data sets I through IV, each consisting of eleven (x, y) points.
- **Carefully study the table**. How do x and y as well as their relationship compare across datasets?
- It shows that all the data sets have nearly identical simple descriptive statistics in terms of mean, standard deviation, correlation, and linear fit!
- **Plotting the data reveals wildly different distributions**, countering the impression that "numerical calculations are exact, but graphs are rough" (Anscombe 1973).
- Graphs 3, Tables 0. Case closed.[1]



[1] That being said, there is a case to be made for tables under certain circumstances. But even tables can (and should) be seen as another form for visualization with clear design principles. To design appealing tables with R, check out the `gt` package together with `gtExtras`.

Data visualization as a method

Different goals, different looks

Exploratory visualization

- "Analytic plots"
- Mostly for ourselves
- Often quick and dirty

Goals

- What's in the data?
- Get a sense of size and complexity of data.
- Explore and interact.
- "Forces us to notice what we never expected to see" (Tukey 1977)

Explanatory visualization

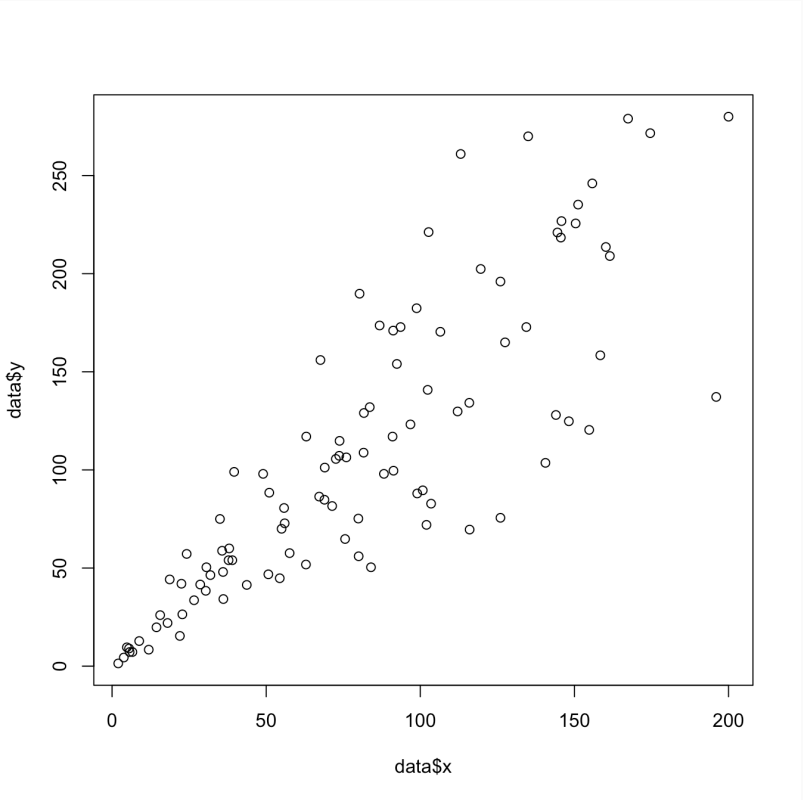
- "Presentation plots"
- For others after the research is completed
- Few, carefully crafted, attractive graphs

Goals

- Communicate content of data.
- Tell a story with data.
- Attract attention and interest.
- "Forces readers to see the information the designer wanted to convey" (Kosslyn 1994)

Different goals, different looks (cont.)

Exploratory visualization



Explanatory visualization

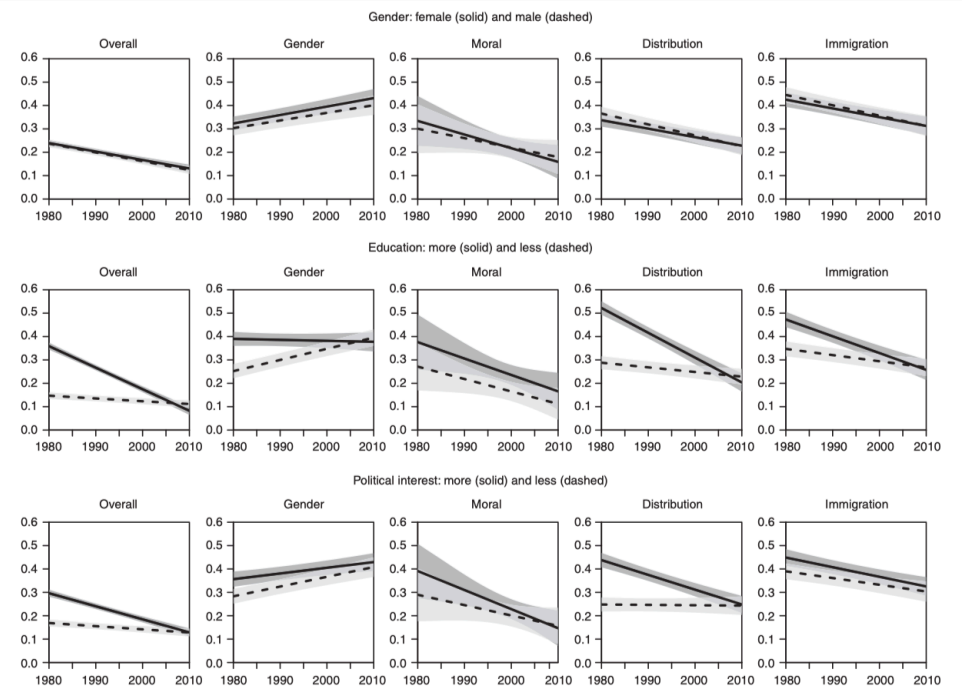


Fig. 4. Polarization trends among several sub-groups
Note: shaded areas around the effects (solid and dashed lines) represent 90 per cent confidence intervals based on simulated responses from the model as a visualization of uncertainty.

The handcraft of visualization

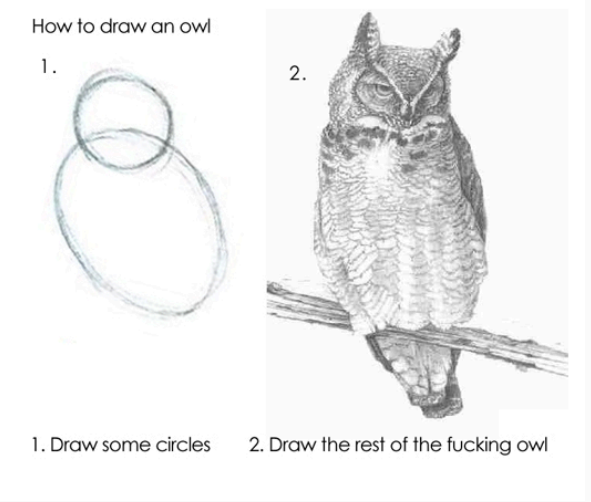
A tool for comparison

- "The fundamental analytical act in statistical reasoning is to answer the question 'compared to what?'
- Whether we are evaluating changes over space or time, searching big data bases, adjusting and controlling for variables, designing experiments, specifying multiple regressions, or doing just about any kind of evidence-based reasoning, the essential point is to make intelligent and appropriate comparisons.
- Thus visual displays, if they are to assist thinking, should show comparisons."

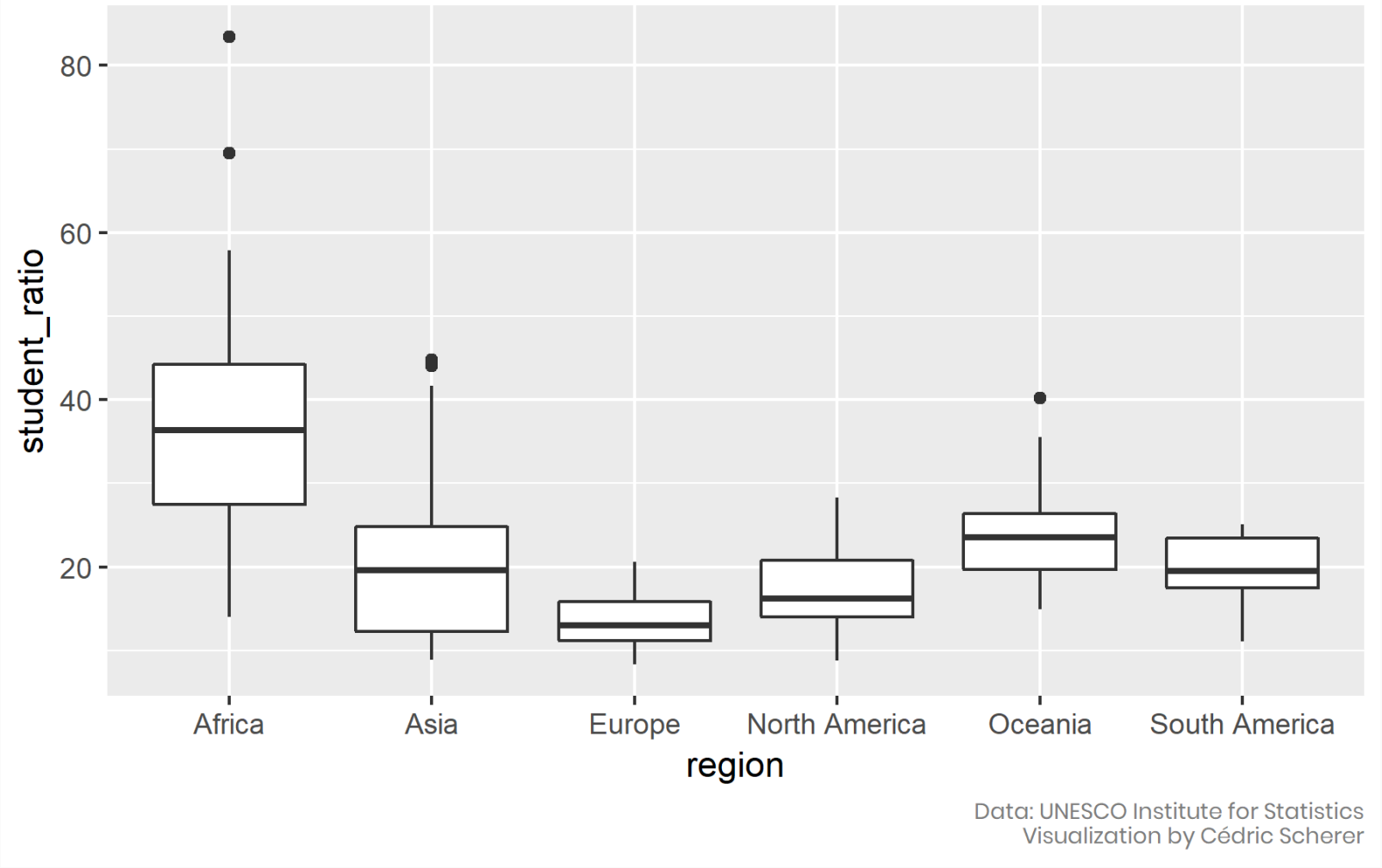
Edward Tufte, Beautiful Evidence, p.127.

Visualization as an iterative process

- The choice of the *right graphical format ultimately depends on the task* or problem it is trying to solve.
- Always *try different graphical formats* on the same data – they may reveal different aspects.
- Constructing visualizations is almost always an **iterative process** – the first graph is rarely also the final one.



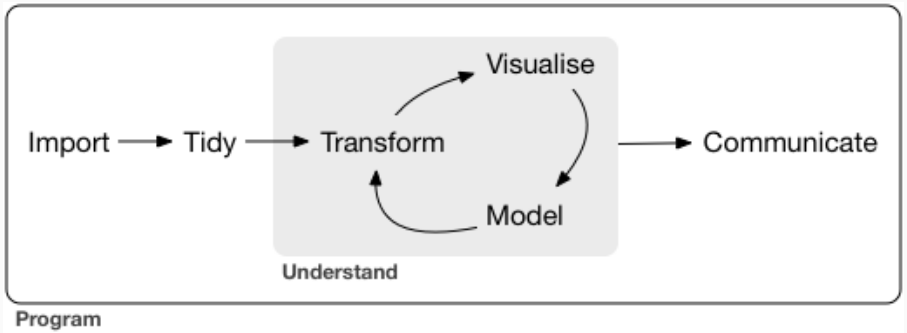
Plotting as an iterative process



`Credit and code` Cedric Scherer

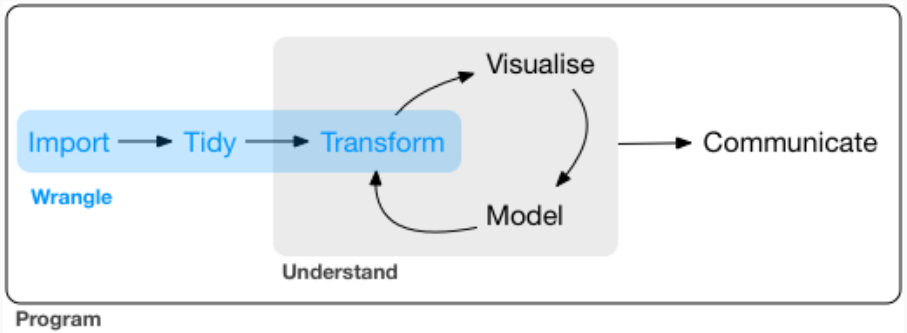
Visualization in the data science workflow

Data visualization is a key skill for communicating insights from data. It is relevant in every step of the workflow.



Visualization in the data science workflow

Data visualization is a key skill for communicating insights from data. It is relevant in every step of the workflow.

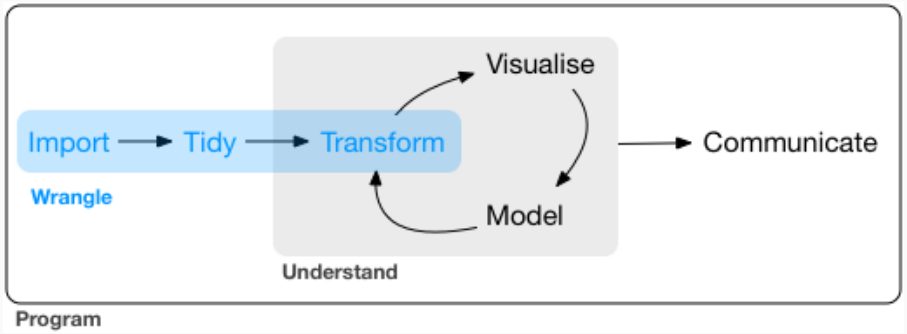


Wrangle

- Sanity checks
- Identification of outliers
- Guidance of recoding operations

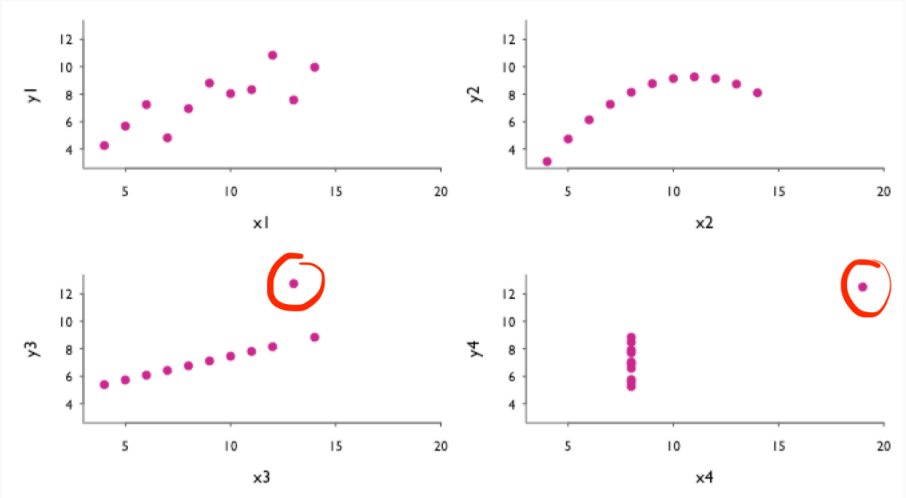
Visualization in the data science workflow

Data visualization is a key skill for communicating insights from data. It is relevant in every step of the workflow.



Wrangle

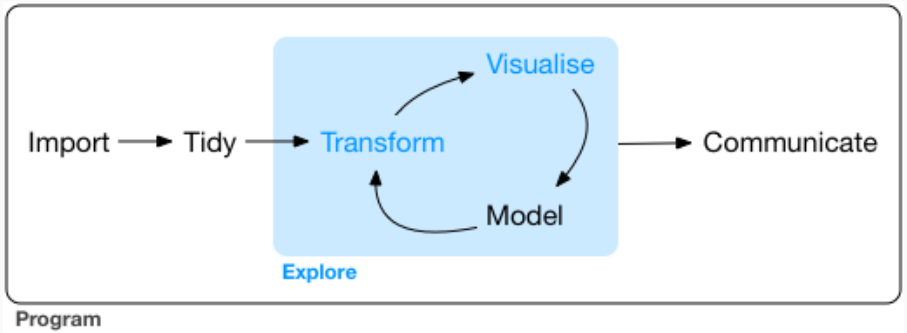
- Sanity checks
- Identification of outliers
- Guidance of recoding operations



Scatter plots to identify outliers in bivariate relationships.

Visualization in the data science workflow

Data visualization is a key skill for communicating insights from data. It is relevant in every step of the workflow.



Wrangle

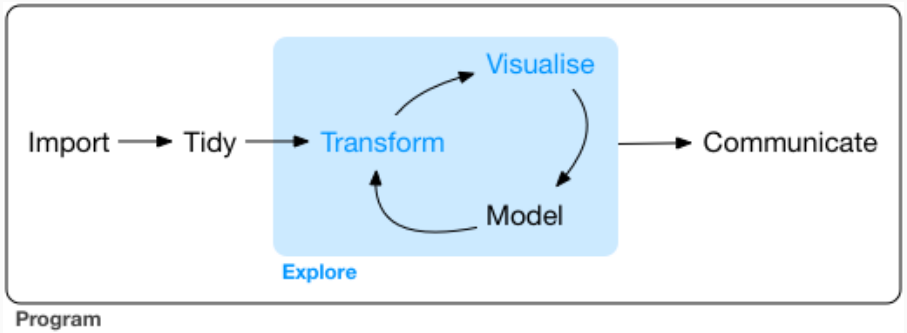
- Sanity checks
- Identification of outliers
- Guidance of recoding operations

Explore

- Summarize distributions
- Discover patterns, relationships

Visualization in the data science workflow

Data visualization is a key skill for communicating insights from data. It is relevant in every step of the workflow.

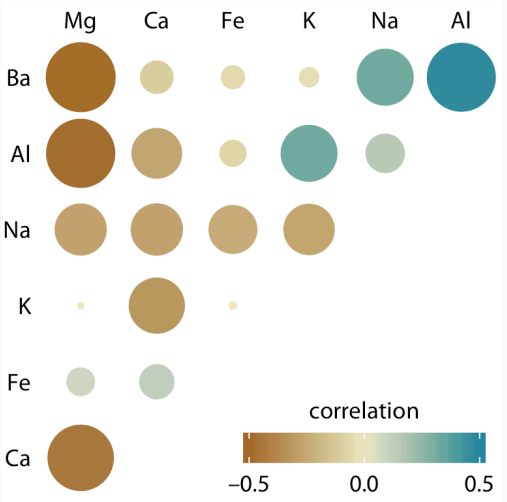


Wrangle

- Sanity checks
- Identification of outliers
- Guidance of recoding operations

Explore

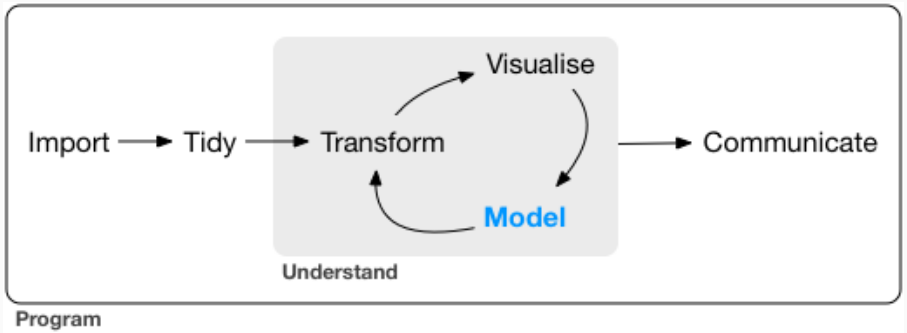
- Summarize distributions
- Discover patterns, relationships



Correlogram to visualize amount of association between pairs of variables

Visualization in the data science workflow (cont.)

Data visualization is a key skill for communicating insights from data. It is relevant in every step of the workflow.

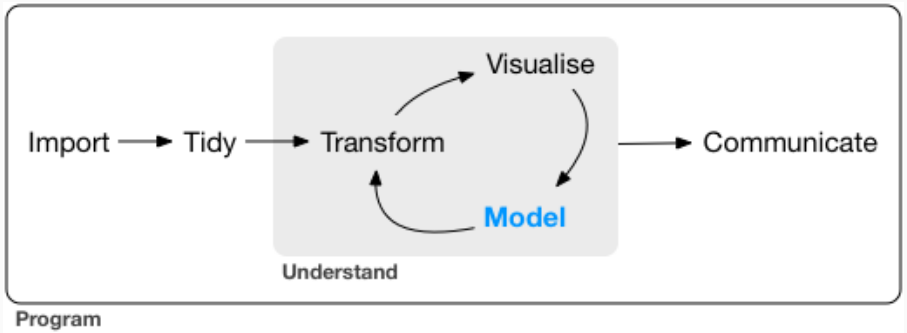


Model

- Test hypotheses
- Summarize (multiple) model estimates
- Visualize uncertainty
- Report robustness/sensitivity analyses

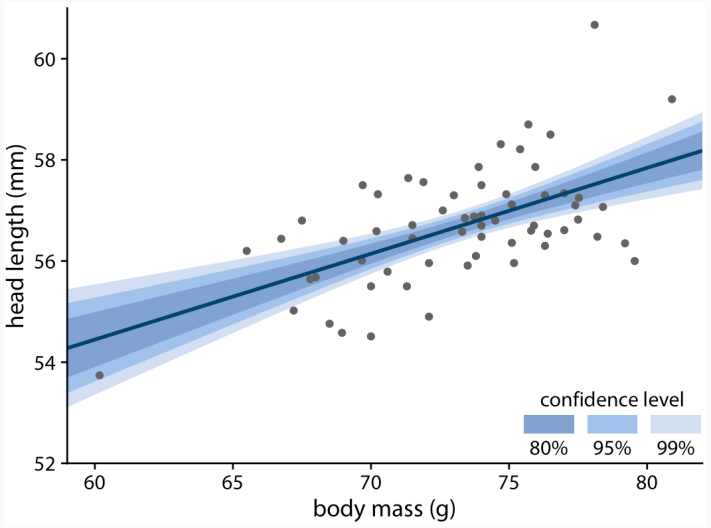
Visualization in the data science workflow (cont.)

Data visualization is a key skill for communicating insights from data. It is relevant in every step of the workflow.



Model

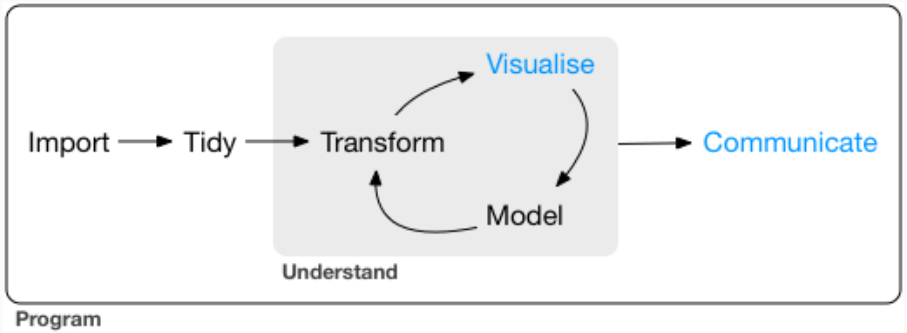
- Test hypotheses
- Summarize (multiple) model estimates
- Visualize uncertainty
- Report robustness/sensitivity analyses



Raw data and trend line with confidence bands to visualize uncertainty of fit

Visualization in the data science workflow (cont.)

Data visualization is a key skill for communicating insights from data. It is relevant in every step of the workflow.



Model

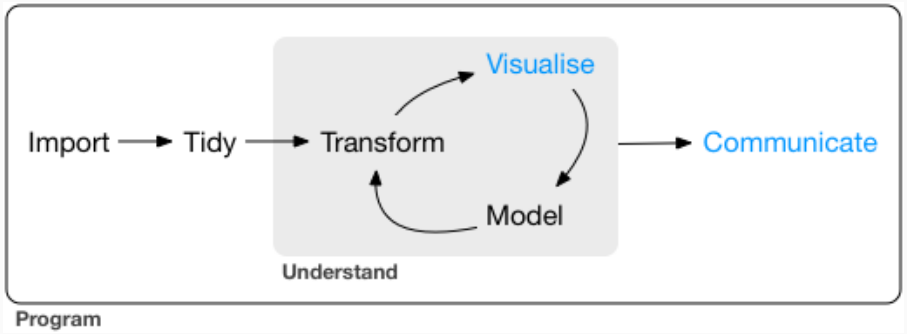
- Test hypotheses
- Summarize (multiple) model estimates
- Visualize uncertainty
- Report robustness/sensitivity analyses

Communicate

- Present raw/cooked data
- Present implications of model results

Visualization in the data science workflow (cont.)

Data visualization is a key skill for communicating insights from data. It is relevant in every step of the workflow.

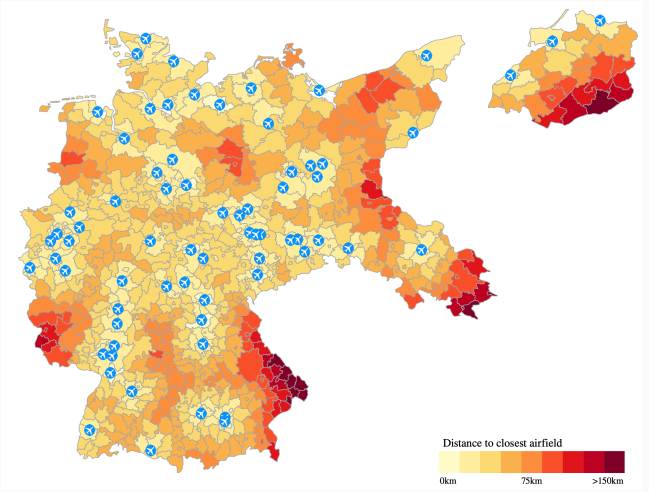


Model

- Test hypotheses
- Summarize (multiple) model estimates
- Visualize uncertainty
- Report robustness/sensitivity analyses

Communicate

- Present raw/cooked data
- Present implications of model results



Choropleth map illustrating the location of civilian airfields in the German Empire, 1932. Administrative counties are shaded according to their centroid's distance to the closest airfield

Visual inference

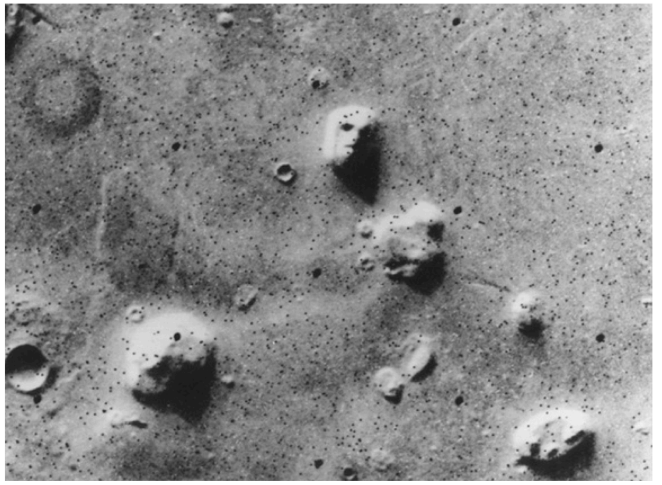
Human talent and weakness

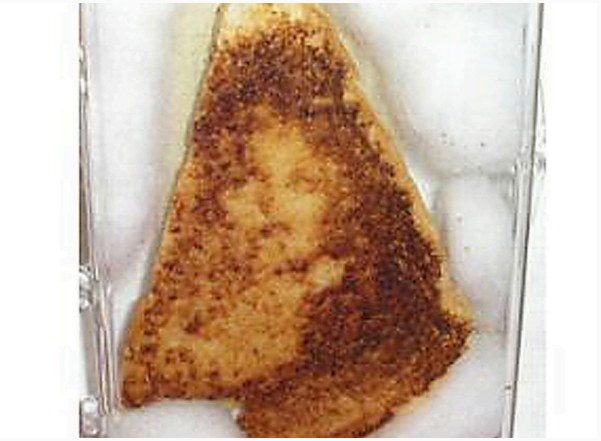
- Humans are extremely good at recognizing patterns.
- At the same time, humans are also extremely good at inferring patterns when there are none (tendency to see patterns in random data = "**apophenia**").
- This is somewhat linked to the fact that our species is bad at dealing with probability and randomness.

Concerns with exploratory data analysis

- A concern that frequently arises with exploratory analysis is that it lacks the rigor of formal tests in confirmatory analysis or conventional statistical inference.
- Long-standing reservations against visualization as merely "informal" approach to data analysis and the fear that beautiful pictures may in fact not correspond to any meaningful patterns of substantive scientific interest.

Image of Mars taken by NASA's Viking 1 orbiter, in grey scale, on July, 25 1976.





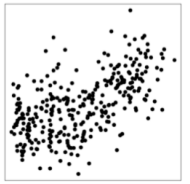
Visual inference (cont.)

Overcoming the exploratory vs. confirmatory visualization divide

- Graphical displays are implicit or explicit comparisons to a reference distribution or baseline model.
- If we discover an interesting pattern in data, this usually means that it looks different from what we expected.
- We usually have implicit models in our mind to which we compare the data ("What do we expect to see?").
- We can make these models explicit and use them to guard against "false discoveries".
- Visual discoveries correspond to the implicit or explicit rejection of null hypotheses (Buja et al. 2009).

Visual inference as an analogue to null hypothesis significance testing

The basic principle of formal testing remains the same in visual inference – with the exception that the test statistic is now a graphical display which is compared to a "reference distribution" of plots showing the null:

Formal Test	Visual Inference
Null hypothesis H_0	Null hypothesis H_0
Test statistic $T = f(x)$	Visual feature in a plot 
Test: Reject? $T(x) > c$?	Human viewer: Discovery?

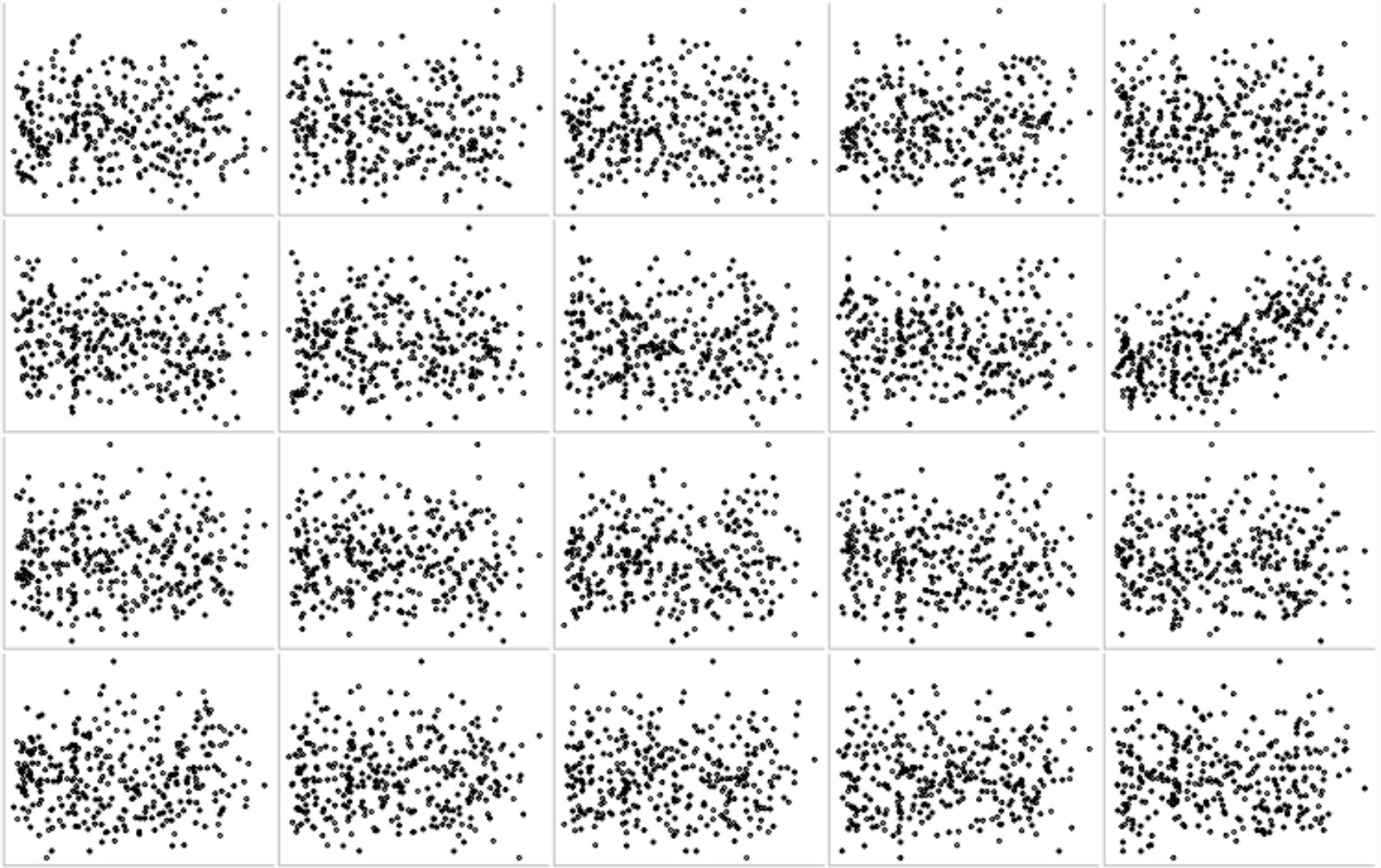
Visual inference: the line-up protocol

This method is called "after the 'police lineup' of criminal investigations [...], because it asks the witness to identify the plot of the real data from among a set of decoys, the null plots, under the veil of ignorance" (Buja et al. 2009).

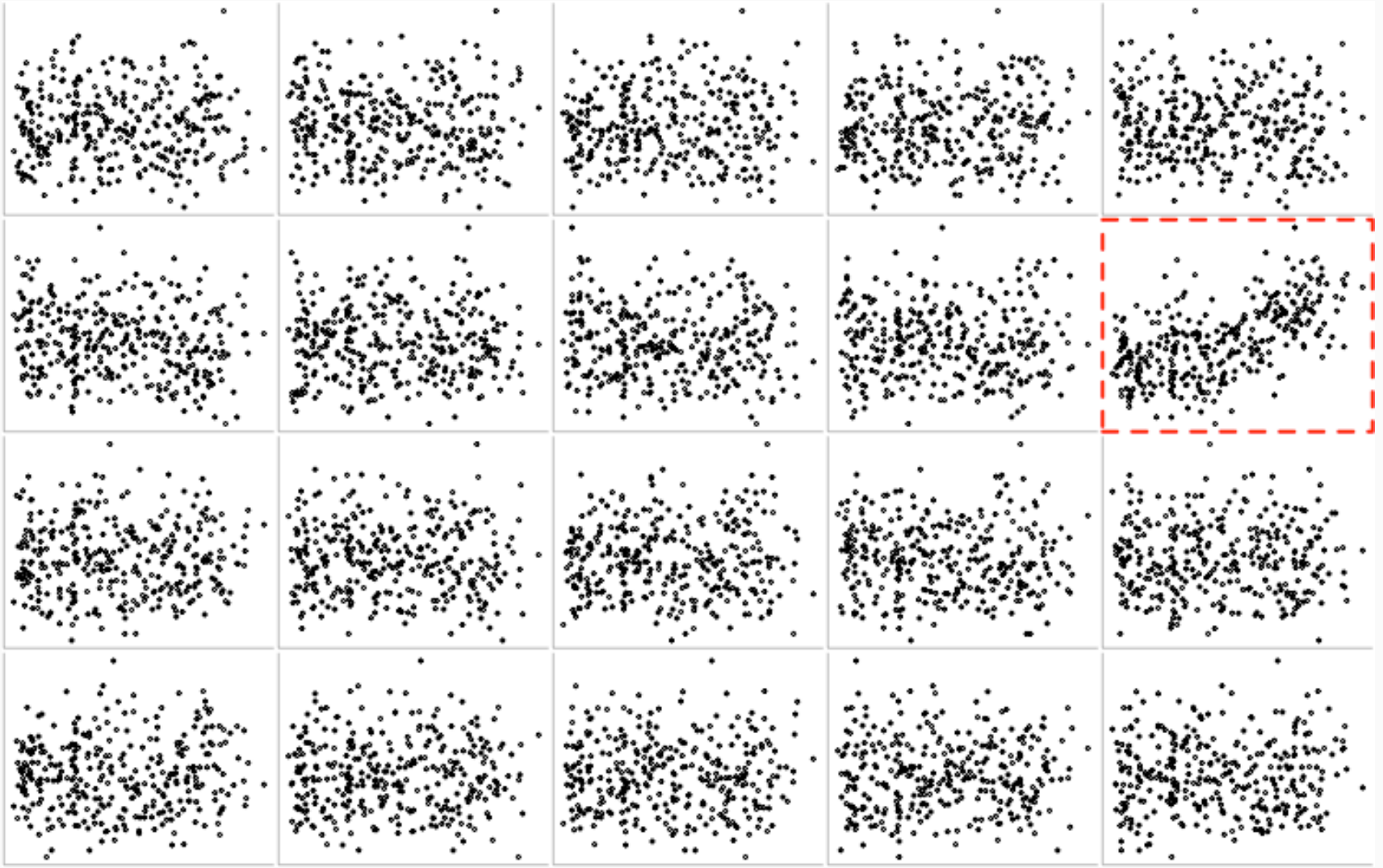
The visual hypothesis test involves the following steps:

1. Simulate data to create $m - 1$ null plots.
2. Randomly place the plot of the real data among them, resulting in a total of m plots.
3. Ask a human viewer to choose the plot that looks the most different from the rest.
4. If the test person succeeds and picks the plot showing the actual data, then this visual discovery can be assigned a p-value of $1/m$. In other words, the probability of picking the true plot just by chance is $1/m$.

Visual inference: which plot stands out?

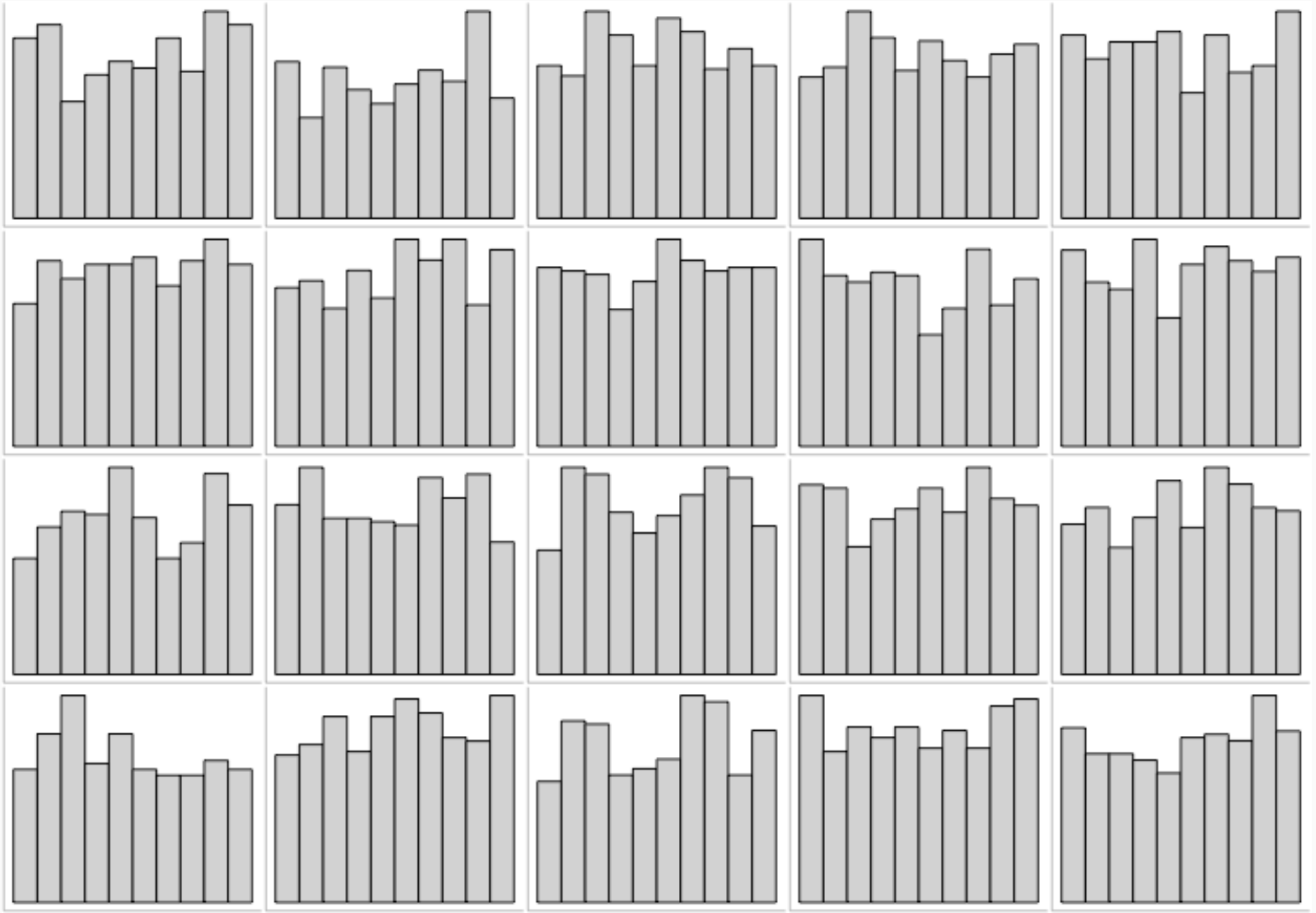


Visual inference: which plot stands out?

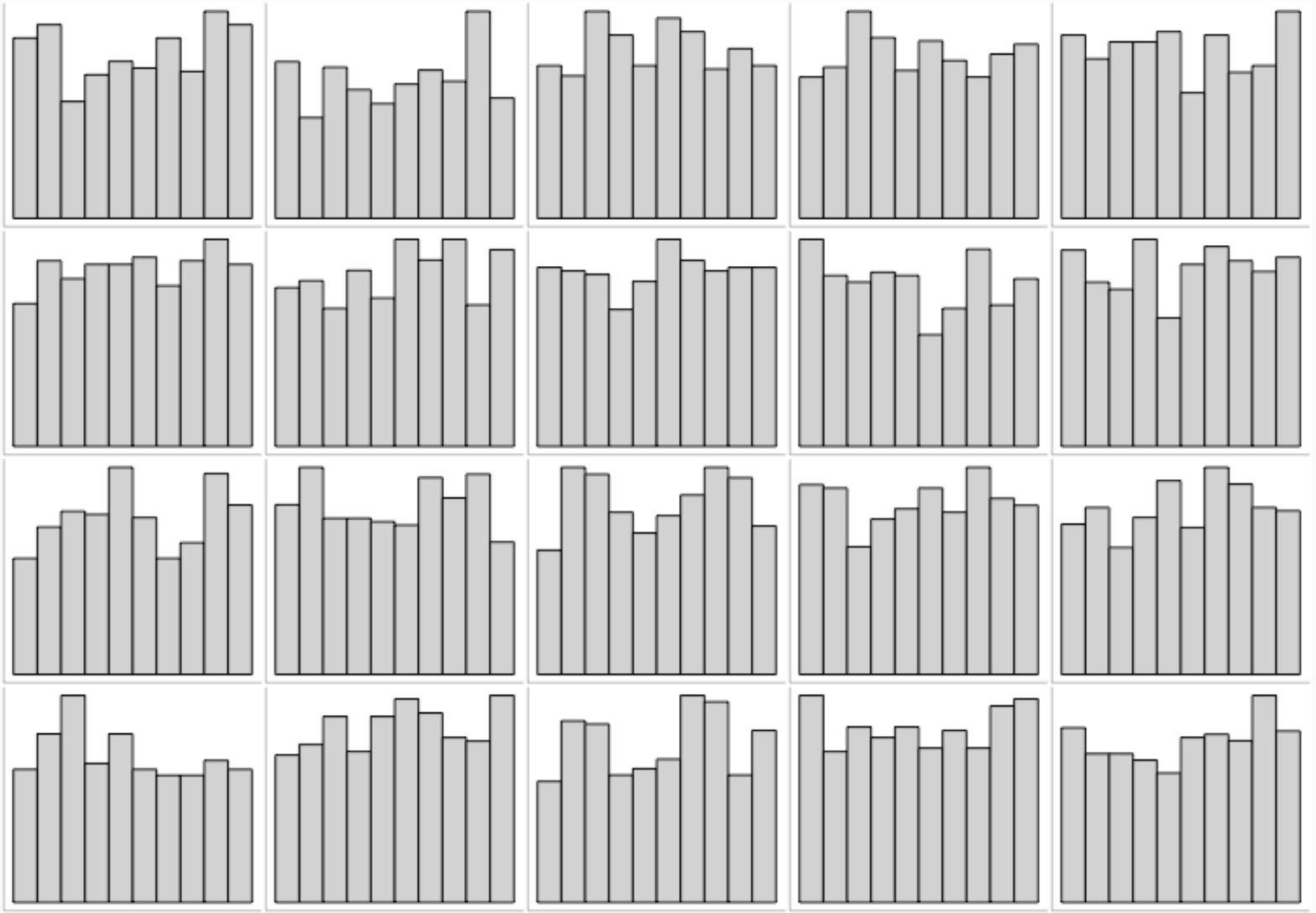


Visual inference: which plot stands out?

Hertie School



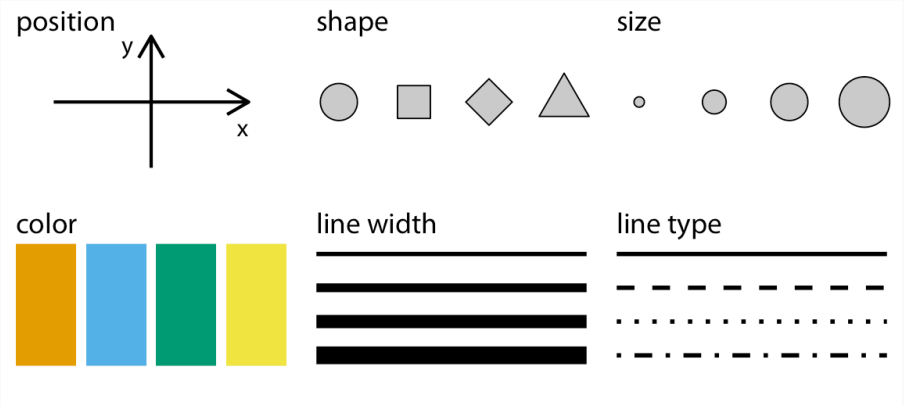
None! All just show random uniform distributions.



Ingredients of data visualization

Mapping data onto aesthetics

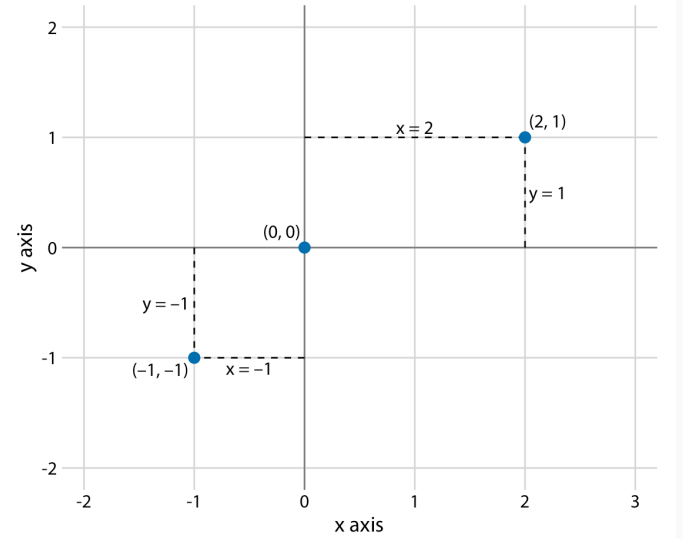
- Whenever we visualize data, we *take data values and convert them* in a systematic and logical way into the visual elements that make up the final graphic.
- Even though there are many *different types of data visualizations*, all these visualizations can be described with a *common language*.
- All data visualizations map data values into **quantifiable features** of the resulting graphic. We refer to these features as **aesthetics**.
- Key aesthetics are:

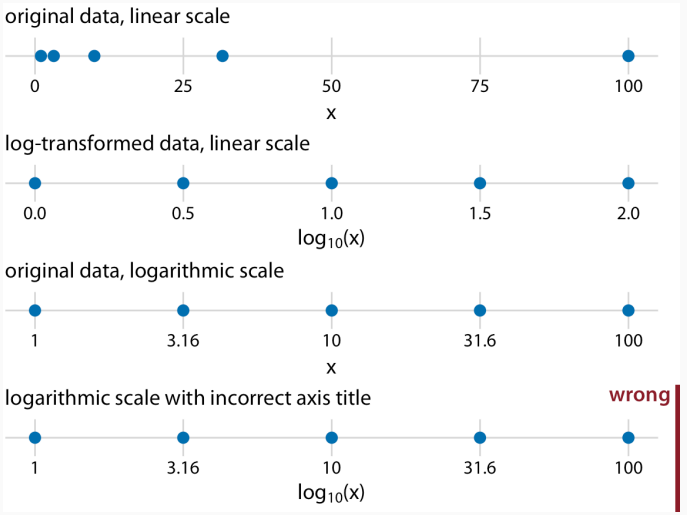


- All **aesthetics fall into one of two groups**: Those that can represent continuous data (e.g., position, size, color) and those that can not (e.g., shape, line type).

Coordinate systems and axes

- **Positions of data values matter**. Usually, we need two position scales (x and y axis of the plot).
- The combination of a set of position scales and their relative geometric arrangement is called a **coordinate system**.
- Often we have two axes representing two different **units**.
- In a **Cartesian** coordinate system, the grid lines along an axis are spaced evenly both in data units and in the resulting visualization.
- There are scenarios where nonlinear scales are preferred. In a nonlinear scale, even spacing in data units corresponds to uneven spacing in the visualization (e.g., log scales).
- Be sure to **label axes properly**!



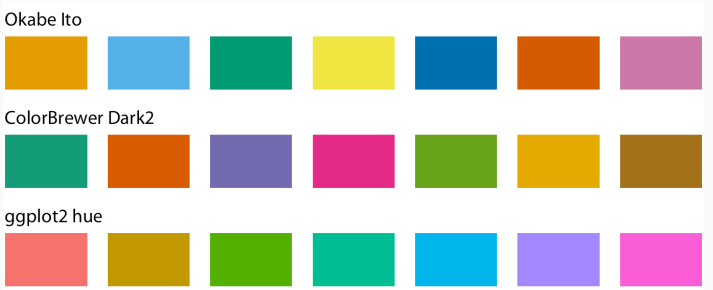


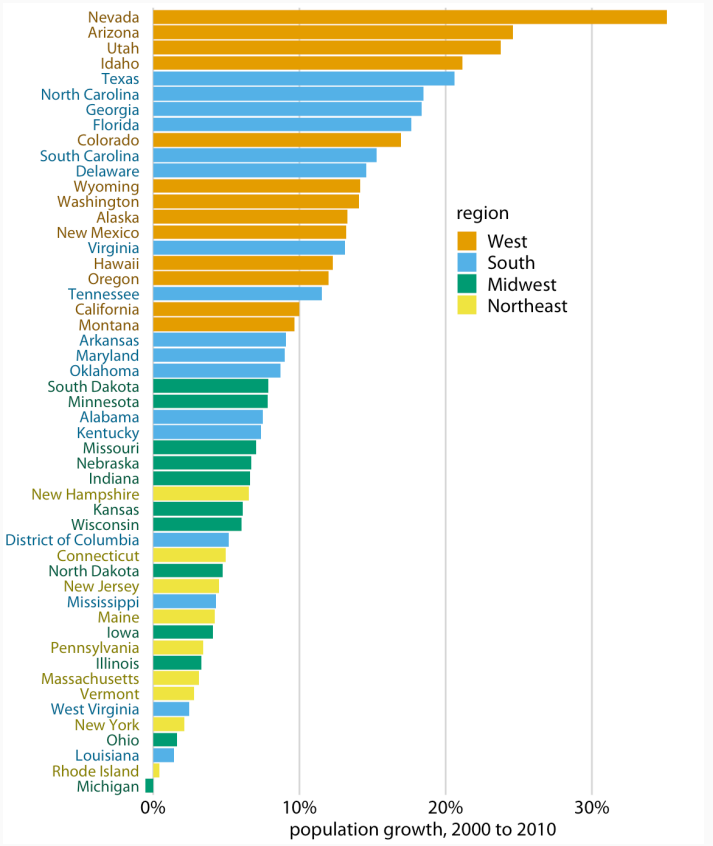
Colors

- There are **three fundamental use cases for color** in data visualizations:
  1. We can use color to distinguish groups of data from each other;
  2. We can use color to represent data values; and
  3. We can use color to highlight.

Colors

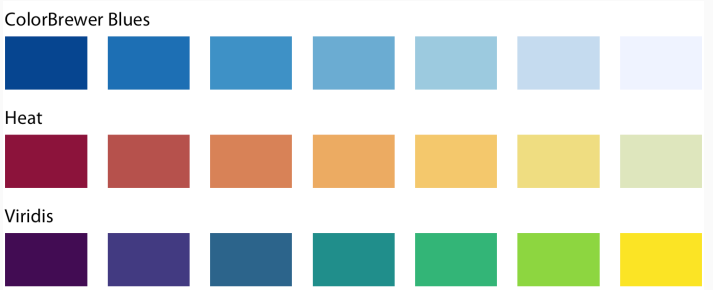
- There are **three fundamental use cases for color** in data visualizations:
  1. We can use color to **distinguish groups of data from each other**;
  2. We can use color to represent data values; and
  3. We can use color to highlight.

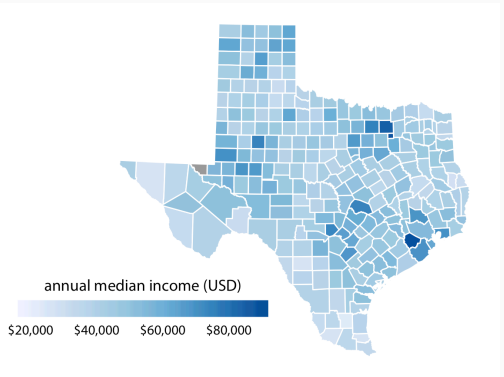


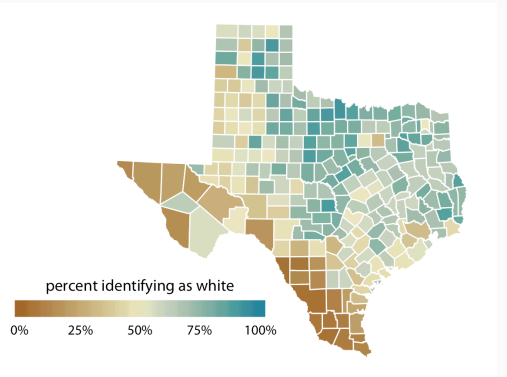


Colors

- There are **three fundamental use cases for color** in data visualizations:
  1. We can use color to distinguish groups of data from each other;
  2. We can use color to **represent data values**; and
  3. We can use color to highlight.

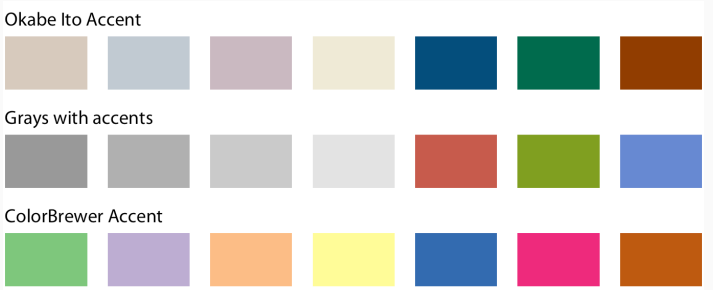


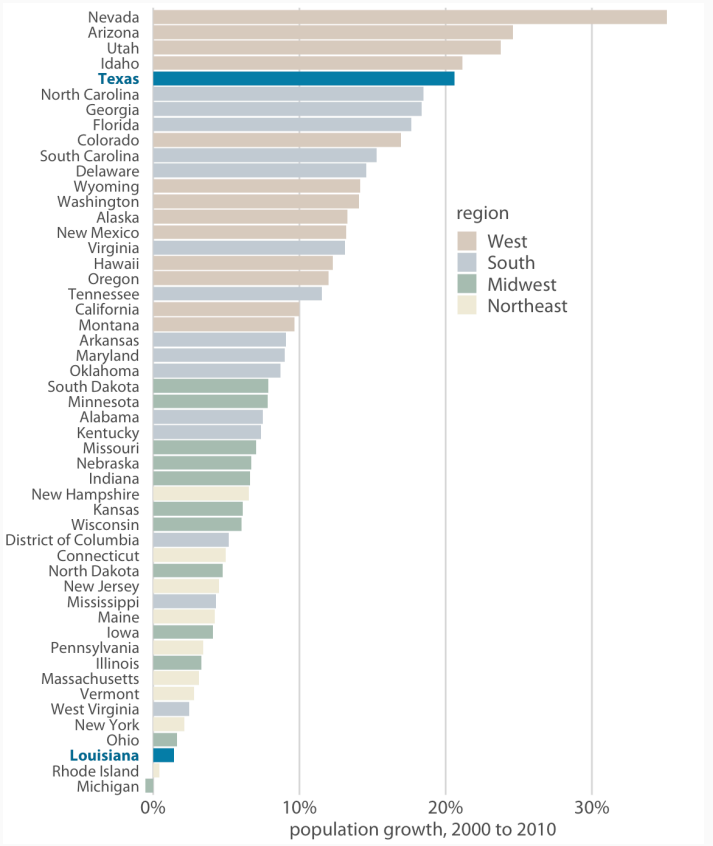




Colors

- There are **three fundamental use cases for color** in data visualizations:
  1. We can use color to distinguish groups of data from each other;
  2. We can use color to represent data values; and
  3. We can use color to **highlight**.



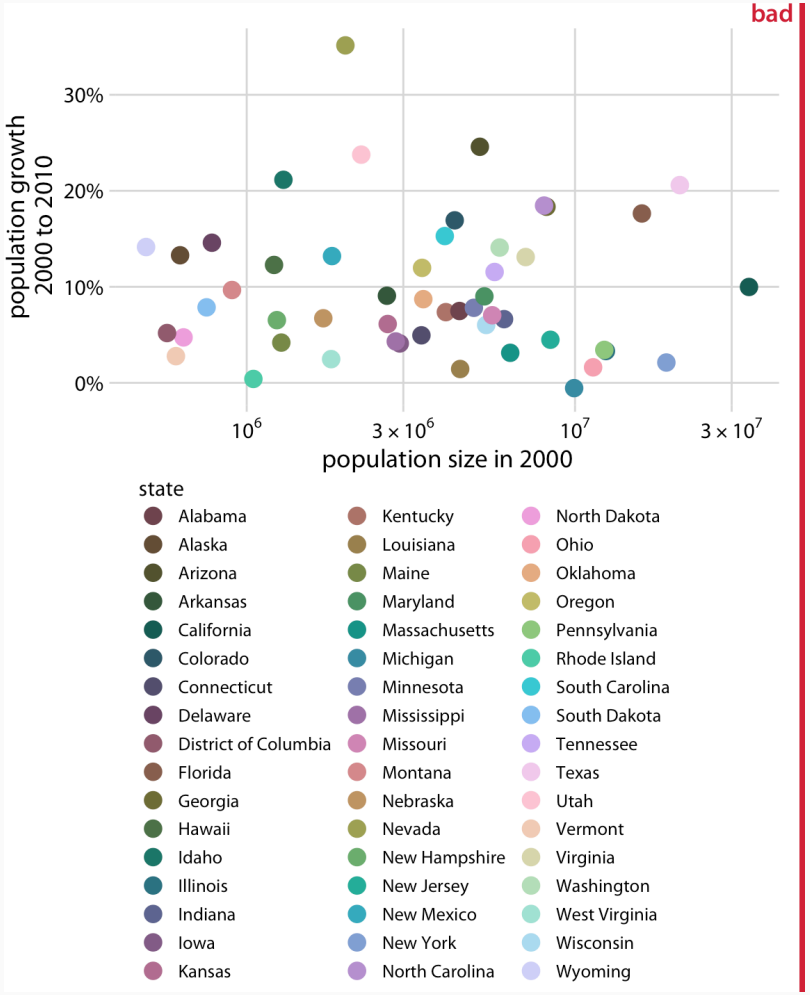


Colors

- There are **three fundamental use cases for color** in data visualizations:
  1. We can use color to distinguish groups of data from each other;
  2. We can use color to represent data values; and
  3. We can use color to highlight.
- While colors are very powerful aesthetics, **try to avoid common pitfalls**, such as:

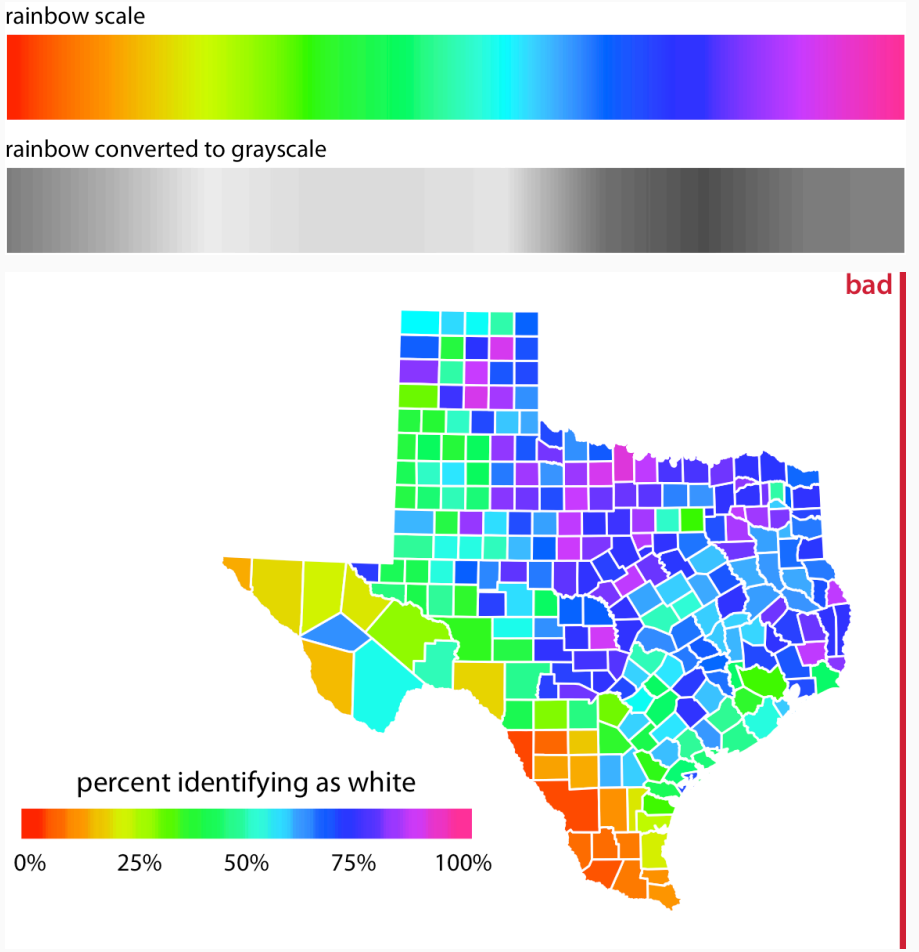
Colors

- There are **three fundamental use cases for color** in data visualizations:
  1. We can use color to distinguish groups of data from each other;
  2. We can use color to represent data values; and
  3. We can use color to highlight.
- While colors are very powerful aesthetics, try to avoid common pitfalls, such as:
  - Encoding **too much / irrelevant information**



Colors

- There are **three fundamental use cases for color** in data visualizations:
  1. We can use color to distinguish groups of data from each other;
  2. We can use color to represent data values; and
  3. We can use color to highlight.
- While colors are very powerful aesthetics, try to avoid common pitfalls, such as:
  - Encoding too much / irrelevant information
  - Using **non-monotonic color scales** to encode data values



- There are **three fundamental use cases for color** in data visualizations:
  1. We can use color to distinguish groups of data from each other;
  2. We can use color to represent data values; and
  3. We can use color to highlight.
- While colors are very powerful aesthetics, try to avoid common pitfalls, such as:
  - Encoding too much / irrelevant information
  - Using non-monotonic color scales to encode data values
  - Not designing for **color-vision deficiency**

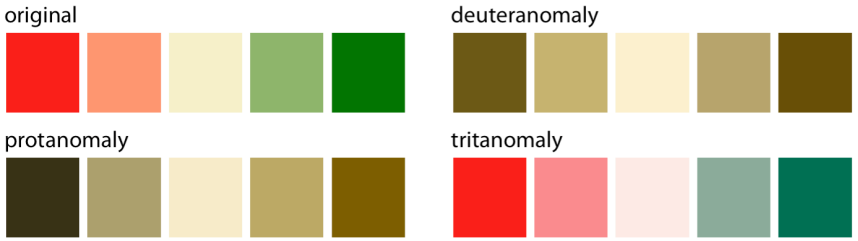


Figure 19.7: A red–green contrast becomes indistinguishable under red–green cvd (deuteranomaly or protanomaly).

Figure 19.8: A blue–green contrast becomes indistinguishable under blue–yellow cvd (tritanomaly).

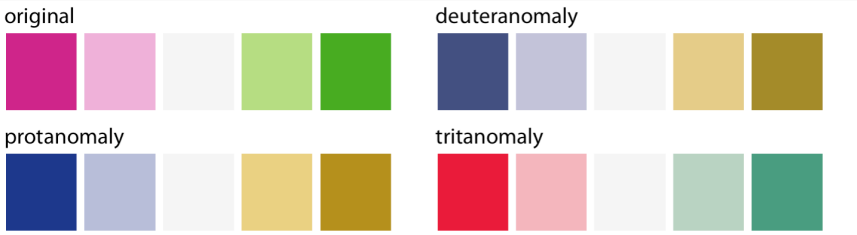


Figure 19.9: The ColorBrewer PiYG (pink to yellow-green) scale from Figure 4.5 looks like a red–green contrast to people with regular color vision but works for all forms of color-vision deficiency. It works because the reddish color is actually pink (a mix of red and blue) while the greenish color also contains yellow. The difference in the blue component between the two colors can be picked up even by deutans or protans, and the difference in the red component can be picked up by tritans.

Colors

- There are **three fundamental use cases for color** in data visualizations:
  1. We can use color to distinguish groups of data from each other;
  2. We can use color to represent data values; and
  3. We can use color to highlight.
- While colors are very powerful aesthetics, try to avoid common pitfalls, such as:
  - Encoding too much / irrelevant information
  - Using non-monotonic color scales to encode data values
  - Not designing for color-vision deficiency
- There's a whole **science around the perception of colors** in graphs, and a range of tools that help you select appropriate color schemes. My favorite is ColorBrewer, which is implemented in the `RColorBrewer` package.

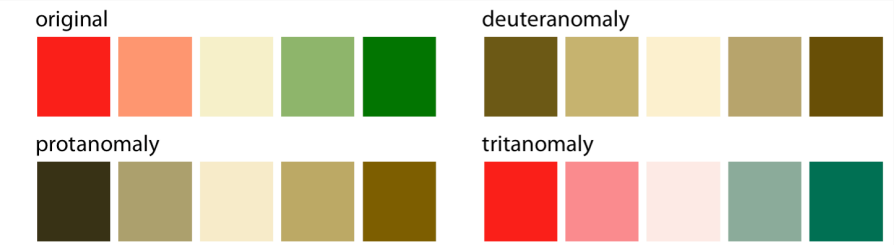


Figure 19.7: A red–green contrast becomes indistinguishable under red–green cvd (deuteranomaly or protanomaly).

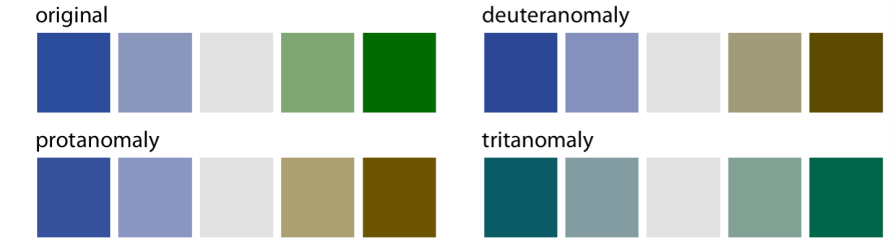


Figure 19.8: A blue–green contrast becomes indistinguishable under blue–yellow cvd (tritanomaly).

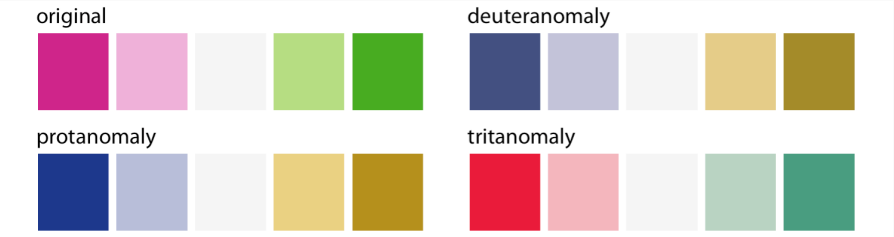


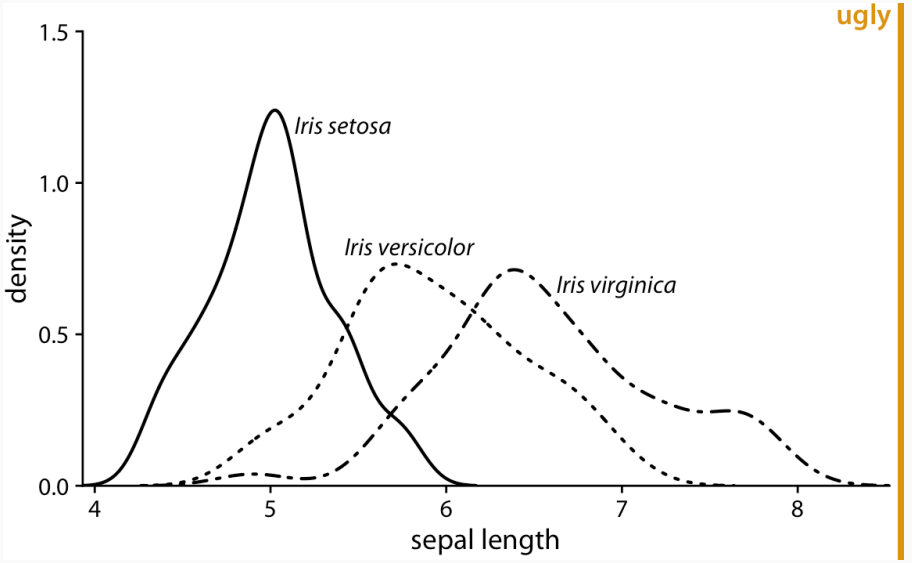
Figure 19.9: The ColorBrewer PiYG (pink to yellow-green) scale from Figure 4.5 looks like a red–green contrast to people with regular color vision but works for all forms of color-vision deficiency. It works because the reddish color is actually pink (a mix of red and blue) while the greenish color also contains yellow. The difference in the blue component between the two colors can be picked up even by deutans or protans, and the difference in the red component can be picked up by tritans.

Line and point types

- Different line and point types **can help distinguish different data types** (e.g., subgroups).
- For lines, we can, use solid, dashed or dotted formatting.
- For points, we can use solid dots, open circles, triangles, or really any symbol we can come up with.
- Try to avoid different line and point types. They add a lot of noise and are difficult to read.

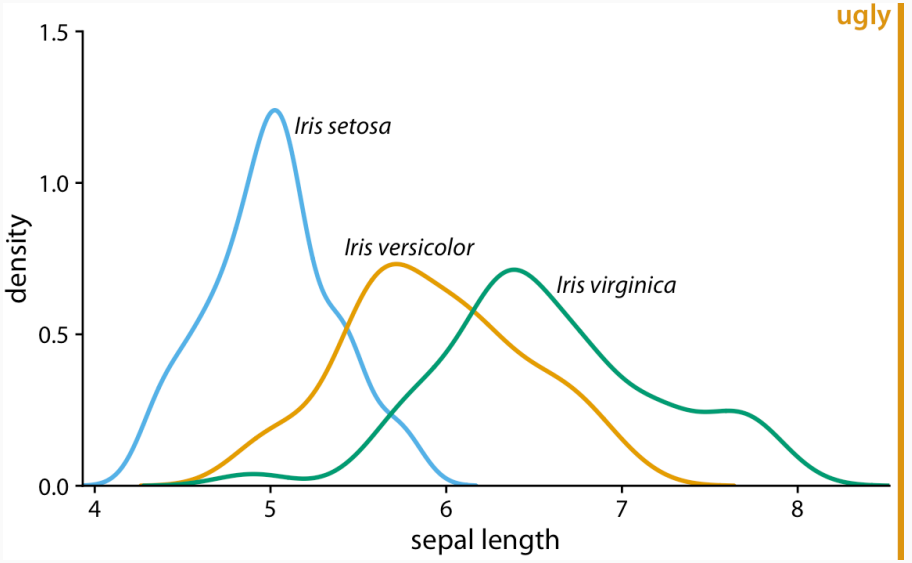
Line and point types

- Different line and point types **can help distinguish different data types** (e.g., subgroups).
- For lines, we can, use solid, dashed or dotted formatting.
- For points, we can use solid dots, open circles, triangles, or really any symbol we can come up with.
- Try to avoid different line and point types. They add a lot of noise and are difficult to read.
- Consider colored shapes instead of **different lines**.



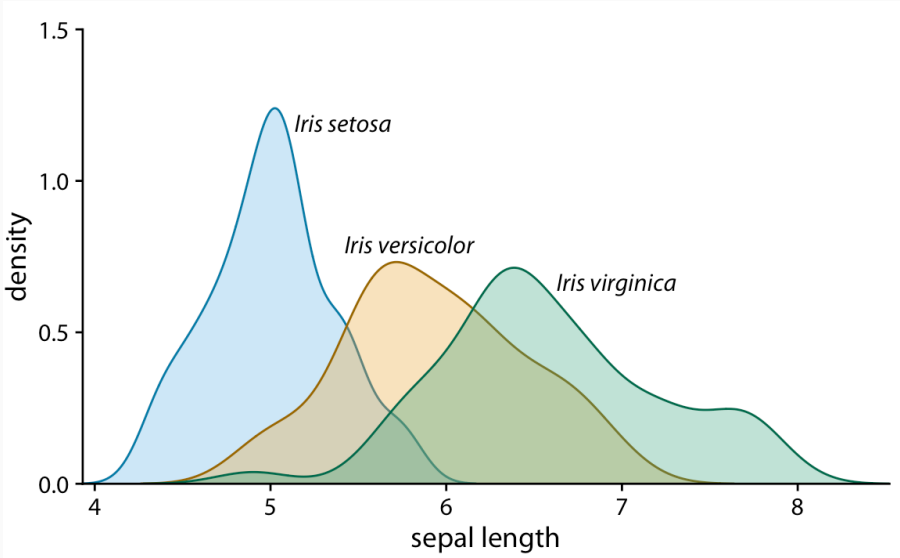
Line and point types

- Different line and point types **can help distinguish different data types** (e.g., subgroups).
- For lines, we can, use solid, dashed or dotted formatting.
- For points, we can use solid dots, open circles, triangles, or really any symbol we can come up with.
- Try to avoid different line and point types. They add a lot of noise and are difficult to read.
- Consider **colored shapes** instead of different lines.



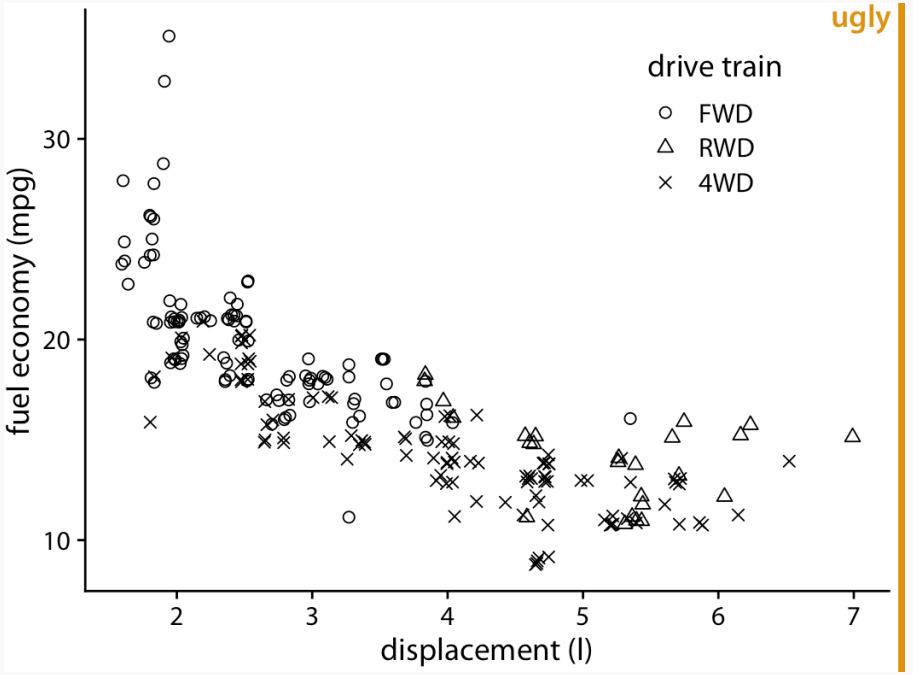
Line and point types

- Different line and point types **can help distinguish different data types** (e.g., subgroups).
- For lines, we can, use solid, dashed or dotted formatting.
- For points, we can use solid dots, open circles, triangles, or really any symbol we can come up with.
- Try to avoid different line and point types. They add a lot of noise and are difficult to read.
- Consider **colored shapes** instead of different lines.



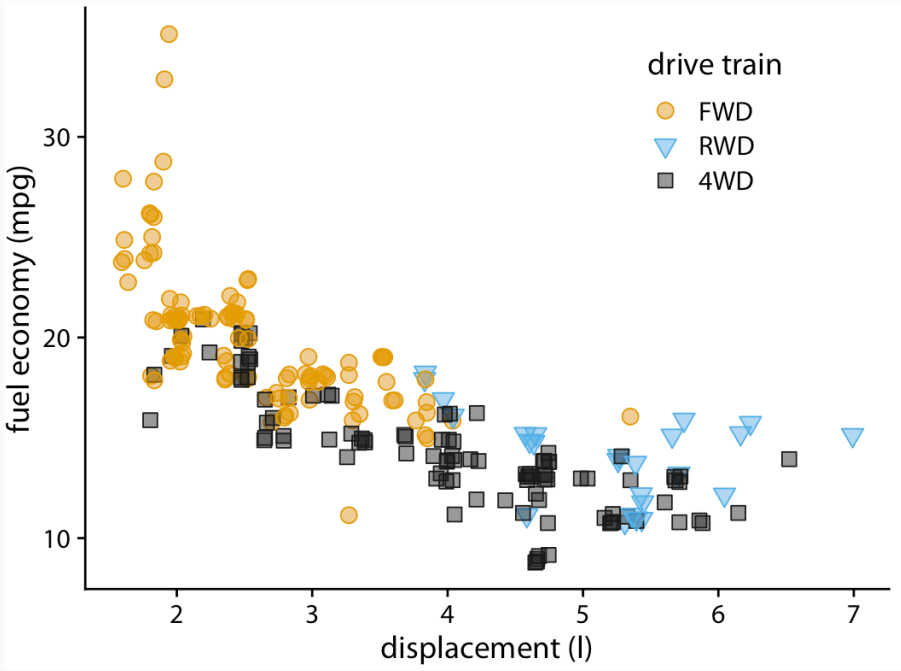
Line and point types

- Different line and point types **can help distinguish different data types** (e.g., subgroups).
- For lines, we can, use solid, dashed or dotted formatting.
- For points, we can use solid dots, open circles, triangles, or really any symbol we can come up with.
- Try to avoid different line and point types. They add a lot of noise and are difficult to read.
- Consider colored shapes instead of different lines.
- Consider colored shapes instead of **different points**.



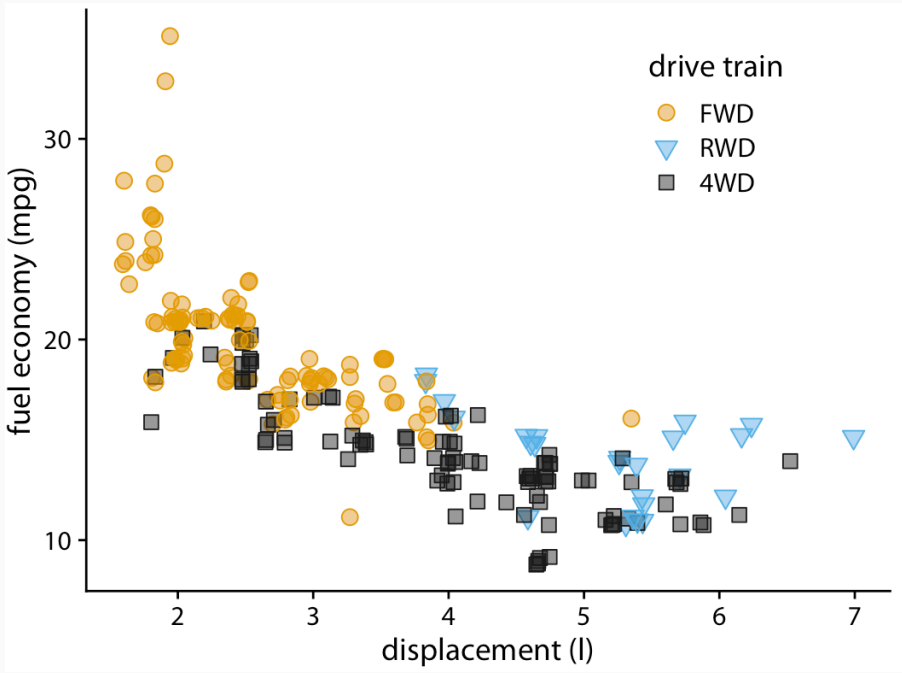
Line and point types

- Different line and point types **can help distinguish different data types** (e.g., subgroups).
- For lines, we can, use solid, dashed or dotted formatting.
- For points, we can use solid dots, open circles, triangles, or really any symbol we can come up with.
- Try to avoid different line and point types. They add a lot of noise and are difficult to read.
- Consider colored shapes instead of different lines.
- Consider **colored shapes** instead of different points.



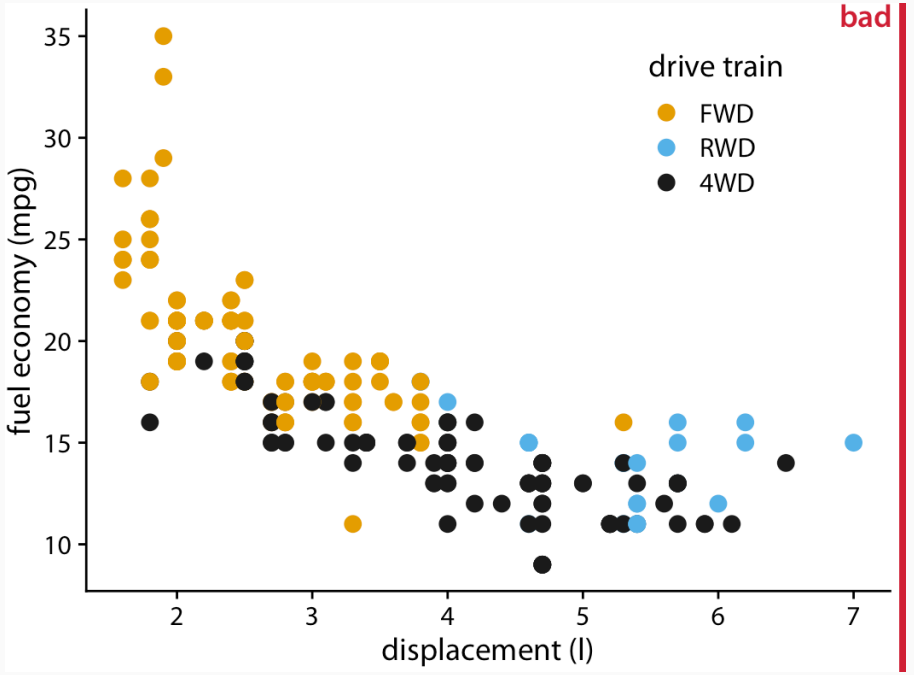
Line and point types

- Different line and point types **can help distinguish different data types** (e.g., subgroups).
- For lines, we can, use solid, dashed or dotted formatting.
- For points, we can use solid dots, open circles, triangles, or really any symbol we can come up with.
- Try to avoid different line and point types. They add a lot of noise and are difficult to read.
- Consider colored shapes instead of different lines.
- Consider colored shapes instead of different points.
- Use **redundant coding**, i.e. use color to enhance the visual appearance of the figure without relying entirely on color to convey key information.



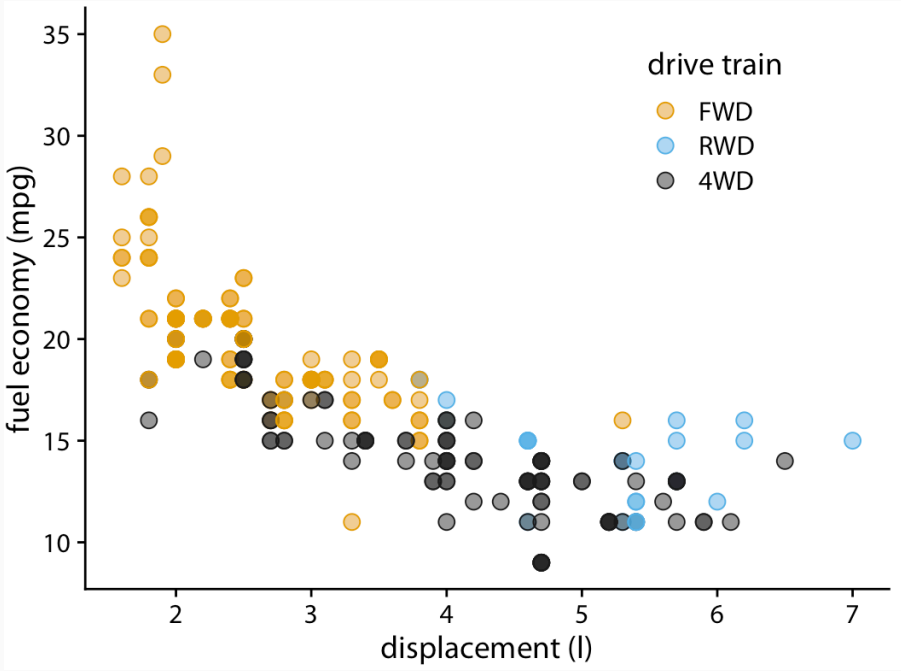
Line and point types

- Different line and point types **can help distinguish different data types** (e.g., subgroups).
- For lines, we can, use solid, dashed or dotted formatting.
- For points, we can use solid dots, open circles, triangles, or really any symbol we can come up with.
- Try to avoid different line and point types. They add a lot of noise and are difficult to read.
- Consider colored shapes instead of different lines.
- Consider colored shapes instead of different points.
- Use redundant coding, i.e. use color to enhance the visual appearance of the figure without relying entirely on color to convey key information.
- To tackle **overlapping data**, use partial transparency (alpha blending) and (moderate) jittering.



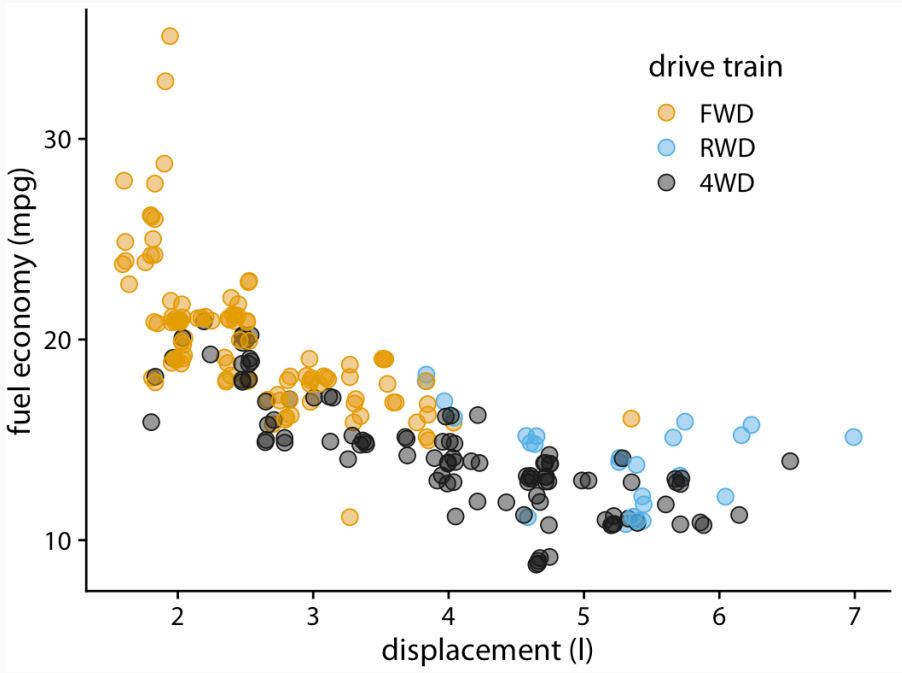
Line and point types

- Different line and point types **can help distinguish different data types** (e.g., subgroups).
- For lines, we can, use solid, dashed or dotted formatting.
- For points, we can use solid dots, open circles, triangles, or really any symbol we can come up with.
- Try to avoid different line and point types. They add a lot of noise and are difficult to read.
- Consider colored shapes instead of different lines.
- Consider colored shapes instead of different points.
- Use redundant coding, i.e. use color to enhance the visual appearance of the figure without relying entirely on color to convey key information.
- To tackle overlapping data, use **partial transparency** (alpha blending) and (moderate) jittering.



Line and point types

- Different line and point types **can help distinguish different data types** (e.g., subgroups).
- For lines, we can, use solid, dashed or dotted formatting.
- For points, we can use solid dots, open circles, triangles, or really any symbol we can come up with.
- Try to avoid different line and point types. They add a lot of noise and are difficult to read.
- Consider colored shapes instead of different lines.
- Consider colored shapes instead of different points.
- Use redundant coding, i.e. use color to enhance the visual appearance of the figure without relying entirely on color to convey key information.
- To tackle overlapping data, use partial transparency (alpha blending) and **(moderate) jittering**.



Types of data visualization

Different plot types for different purposes

A common mistake in visualization is that **plot types are used for purposes they are not meant for**. You'll gain more intuition and experience in picking the right types over time.

Before you start plotting, ask yourself:

1. Which quantity do I want to visualize?

- Amounts
- Distributions
- Proportions
- Associations
- Structures
- Trends
- Estimates
- Predictions
- Uncertainty

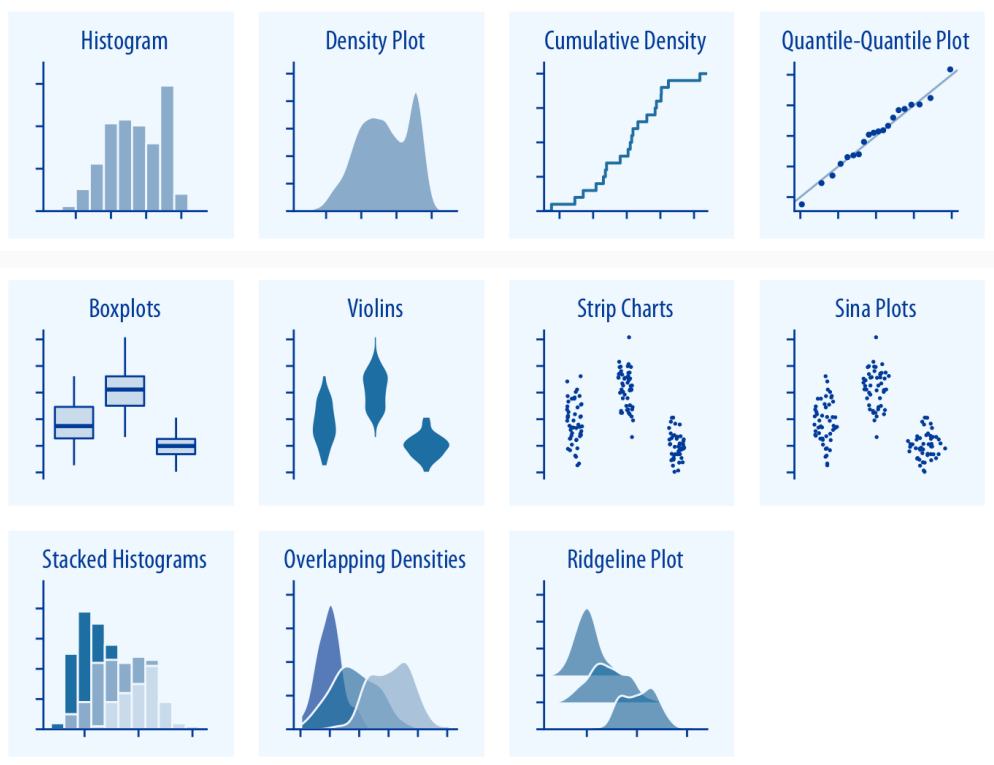
2. Which question do I want to answer?

- "Is the *distribution* normal (or uniform or...)?" → **Histogram, density plot, Q-Q plot**
- "Are univariate *distributions* across subgroups different?" → **Boxplots, ridgelines**
- "How do *differences in amounts* between groups compare?" → **Barplot, dotplot**
- "What is the *relationship* between x and y?" → **Scatterplot, contour plot, hex bins**
- "What are the *correlations* in a set of variables?" → **Correlogram, small multiples**
- "How did a *trend* develop over time?" → **Line graph, slopegraph**
- "Are the data *clustered* by subgroup?" → **Scatterplot with color**
- "Is there a *spatial pattern*?" → **Choropleth, cartogram heatmap**
- "What are the relative and absolute *effect sizes*? → **Coefficient plot**
- "How uncertain are *estimates*? → **Error bars, confidence bands**

A directory of visualizations

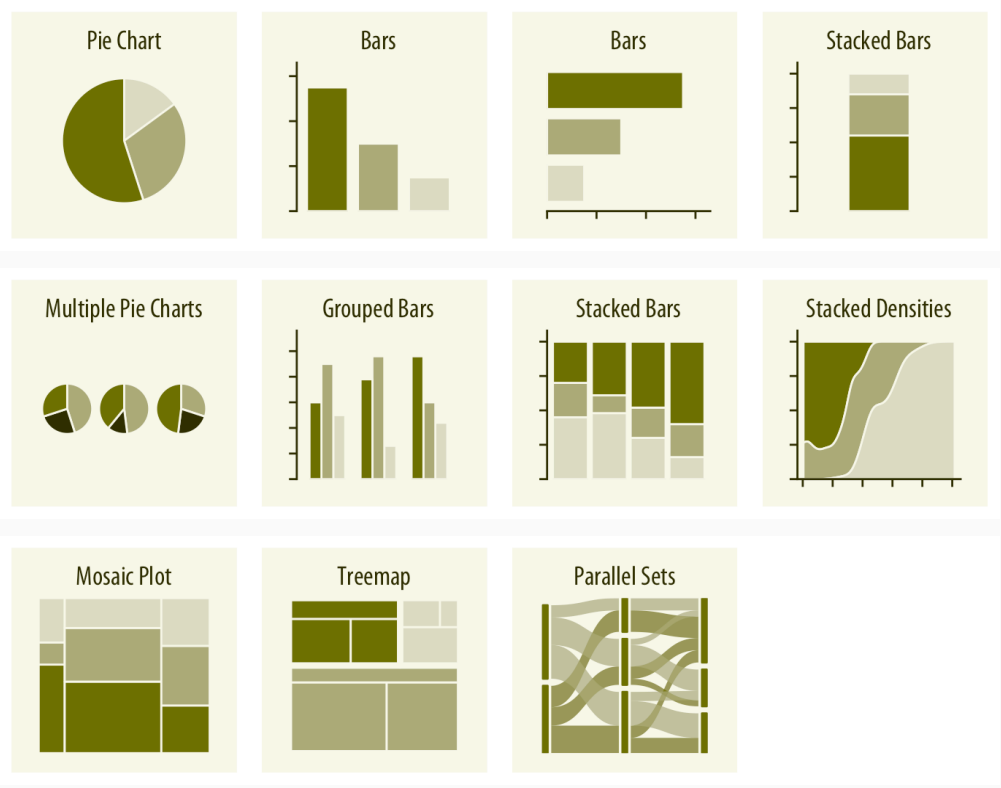
Visualizing amounts

Visualizing distributions

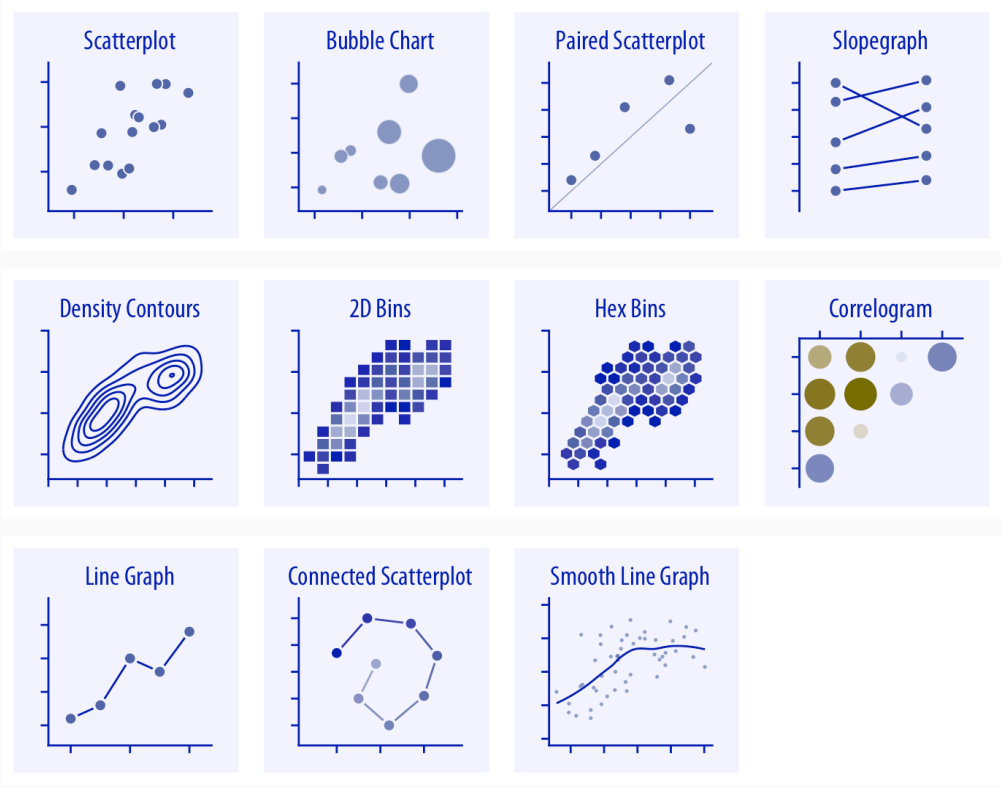


A directory of visualizations (cont.)

Visualizing proportions

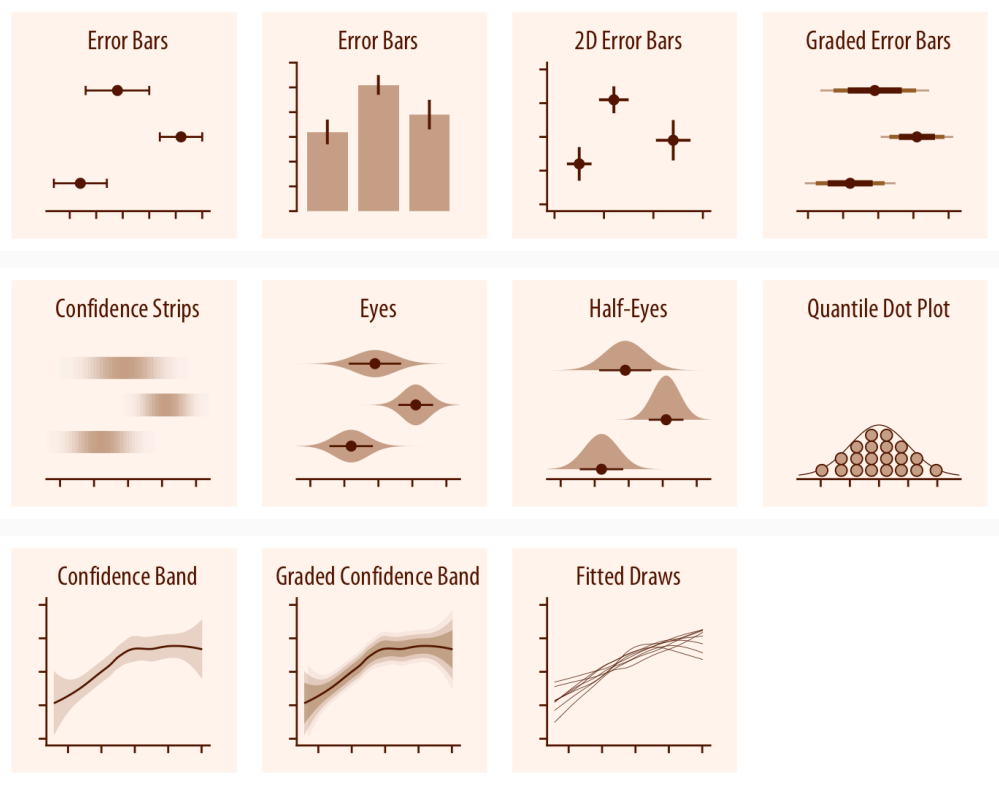


Visualizing x-y relationships

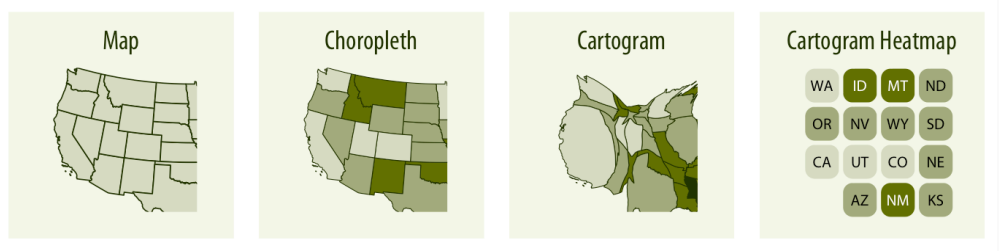


A directory of visualizations (cont.)

Visualizing uncertainty

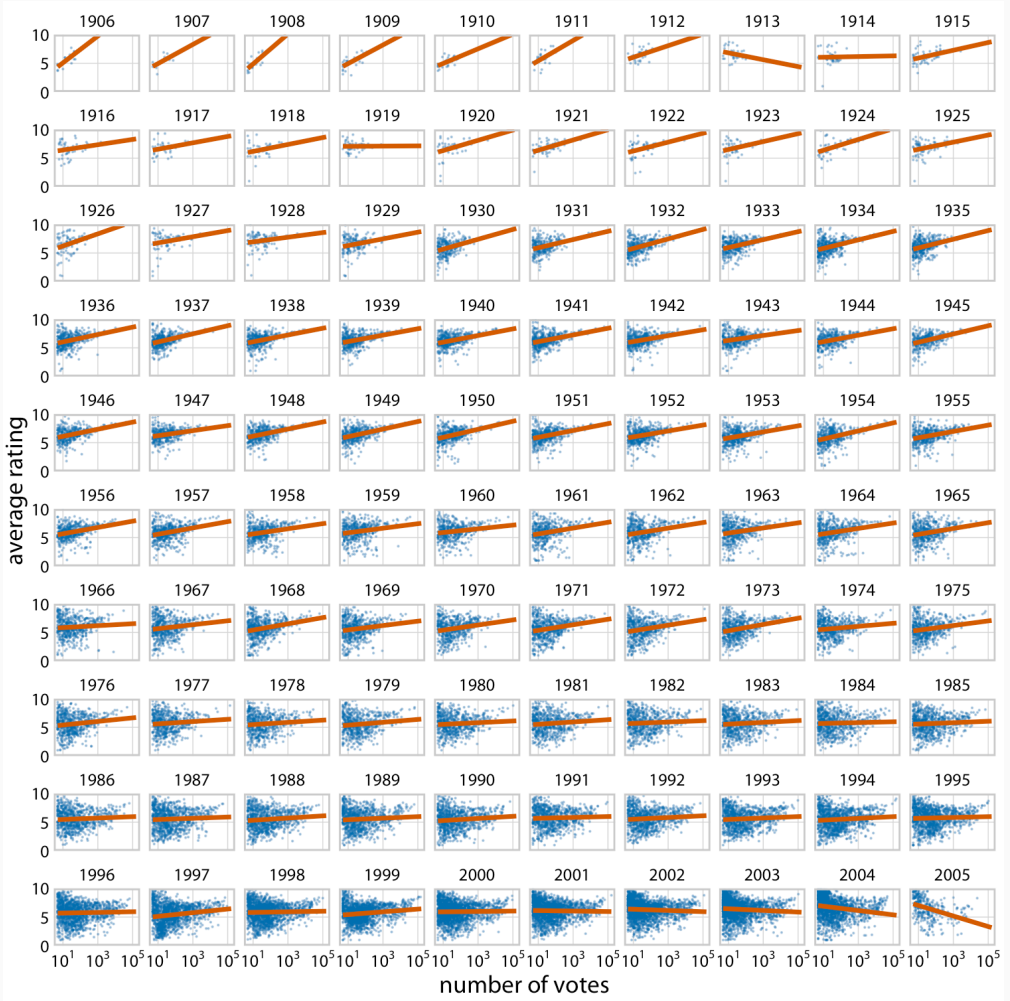


Visualizing geospatial data



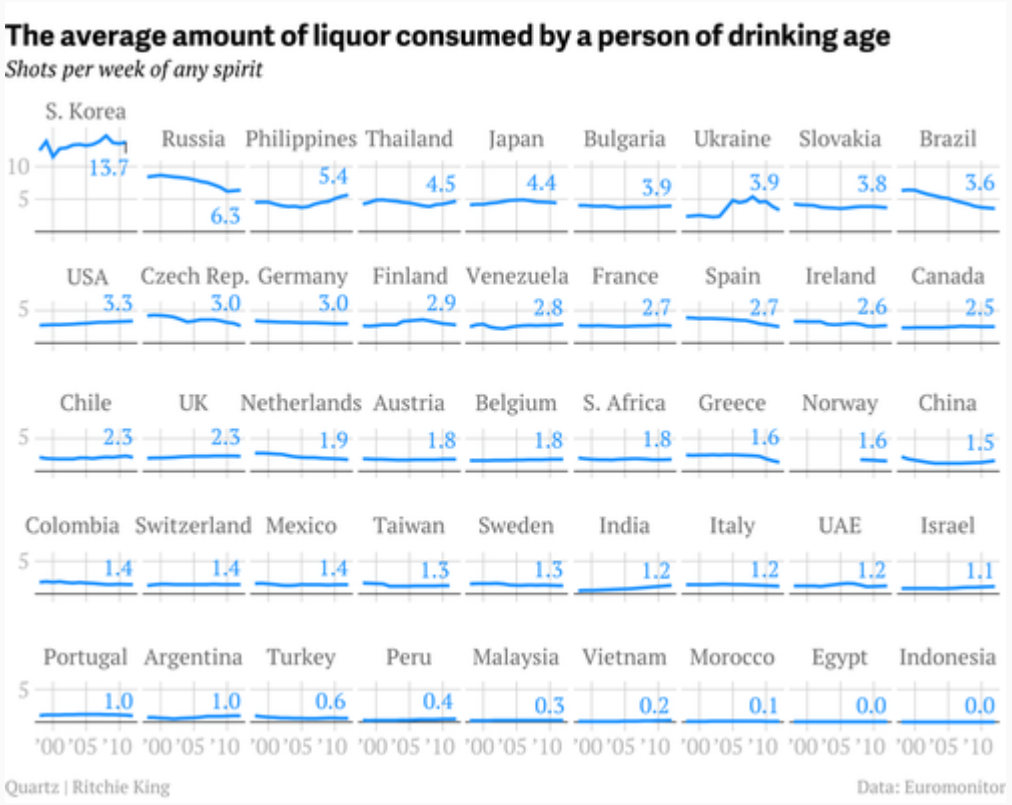
Small multiples

- A powerful yet underestimated visualization strategy is to use multi-panel figures.
- Often we want to compare relationships or trends between groups. With many groups, that's too much information for a single figure panel.
- There are various terms for multi-panel figures, including "small multiples" (Tufte 1990), "trellis plot" (Cleveland 1993) and "faceting" (Wickham 2016)
- In R we can implement this fairly easily with `ggplot`'s `facet_grid()` (or `facet_wrap()`).
- If you do small multiples, make sure to use:
  - common graph size
  - common axis scales
  - helpful alignment and order of panels



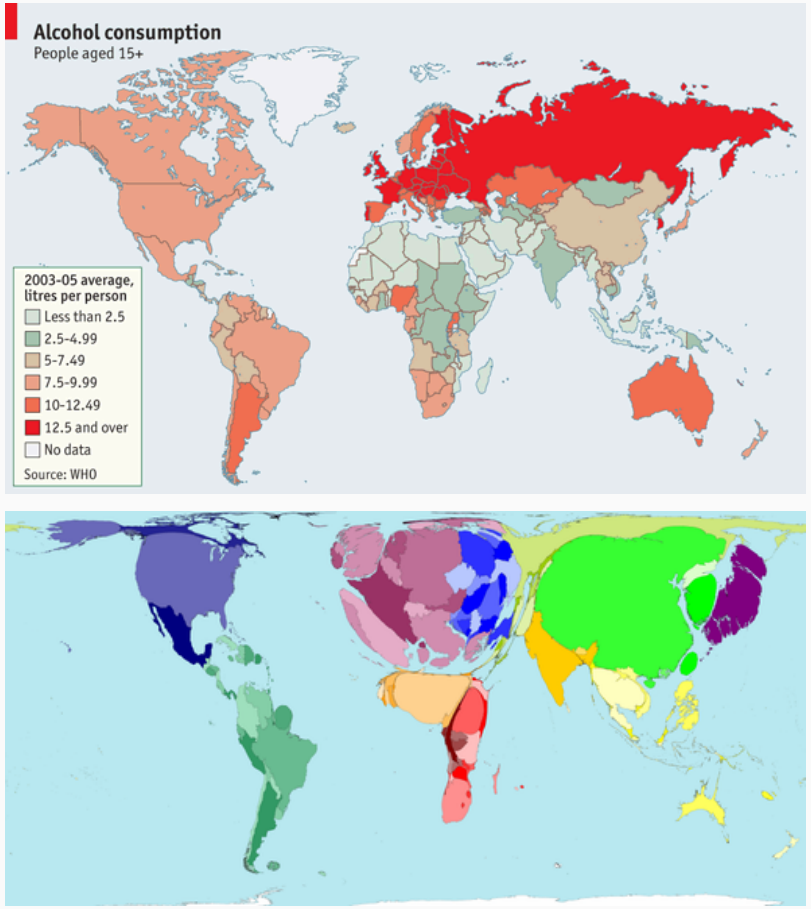
Small multiples (cont.)

- Check out the example on the right, which was also discussed here and there.
- What they did right (by Kaiser Fung):
  - Did not put the data on a map
  - Ordered the countries by the most recent data point rather than alphabetically
  - Scale labels are found only on outer edge of the chart area, rather than one set per panel
  - Only used three labels for the 11 years
  - Did not overdo the vertical scale either
  - The nicest feature was the XL scale applied only to South Korea. This destroys the small-multiples principle but draws attention to the top left corner, where the designer wants our eyes to go.



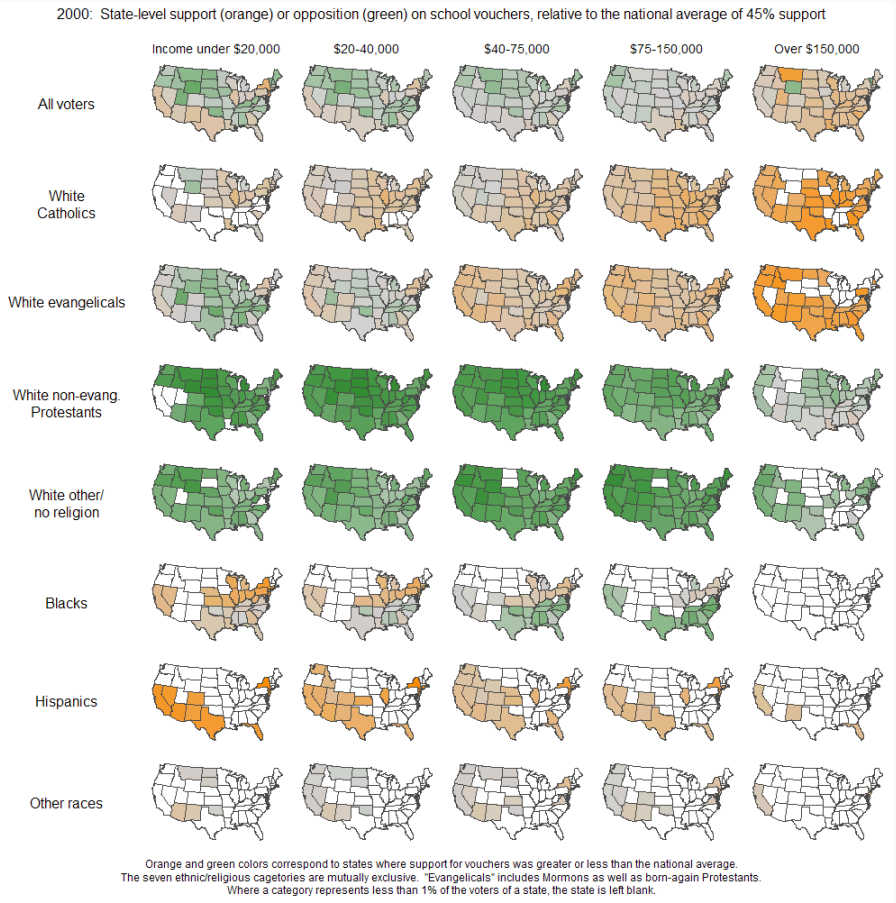
Small multiples (cont.)

- Check out the example on the right, which was also discussed here and there.
- What they did right (by Kaiser Fung):
  - Did not put the data on a map
  - Ordered the countries by the most recent data point rather than alphabetically
  - Scale labels are found only on outer edge of the chart area, rather than one set per panel
  - Only used three labels for the 11 years
  - Did not overdo the vertical scale either
  - The nicest feature was the XL scale applied only to South Korea. This destroys the small-multiples principle but draws attention to the top left corner, where the designer wants our eyes to go.
- Sometimes maps are not a good alternative.



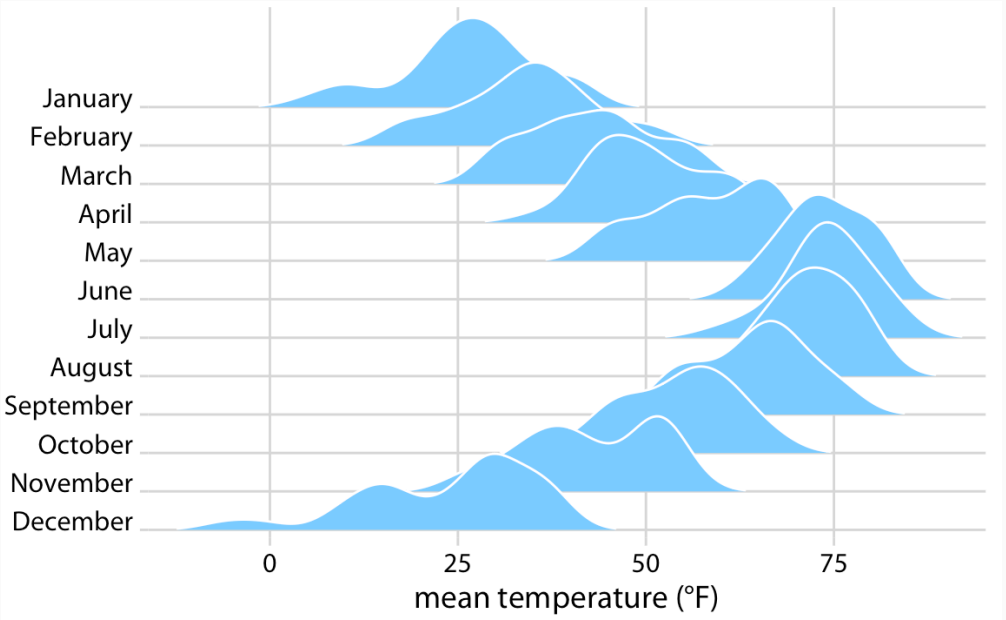
Small multiples (cont.)

- Check out the example on the right, which was also discussed here and there.
- What they did right (by Kaiser Fung):
  - Did not put the data on a map
  - Ordered the countries by the most recent data point rather than alphabetically
  - Scale labels are found only on outer edge of the chart area, rather than one set per panel
  - Only used three labels for the 11 years
  - Did not overdo the vertical scale either
  - The nicest feature was the XL scale applied only to South Korea. This destroys the small-multiples principle but draws attention to the top left corner, where the designer wants our eyes to go.
- Sometimes maps are not a good alternative.
- But you can do small multiples of maps!



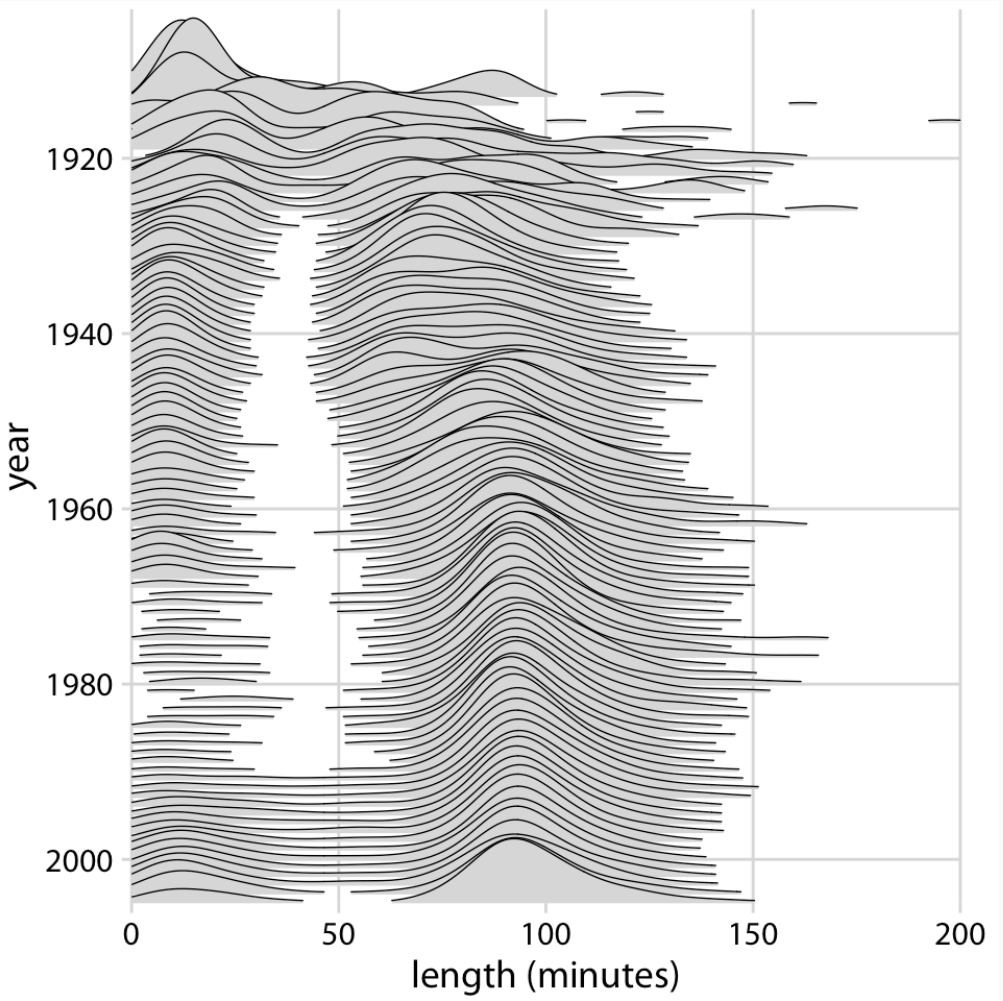
Visualizing many distributions at once

- An increasingly popular, compact variant of small multiples for distributions is the **ridgeline plot** (they look like mountain ridgelines).
- The idea is to staggering distributions plots in the vertical direction (i.e., along the horizontal axis).
- Ridgeline plots tend to work particularly well if want to show trends in distributions over time.



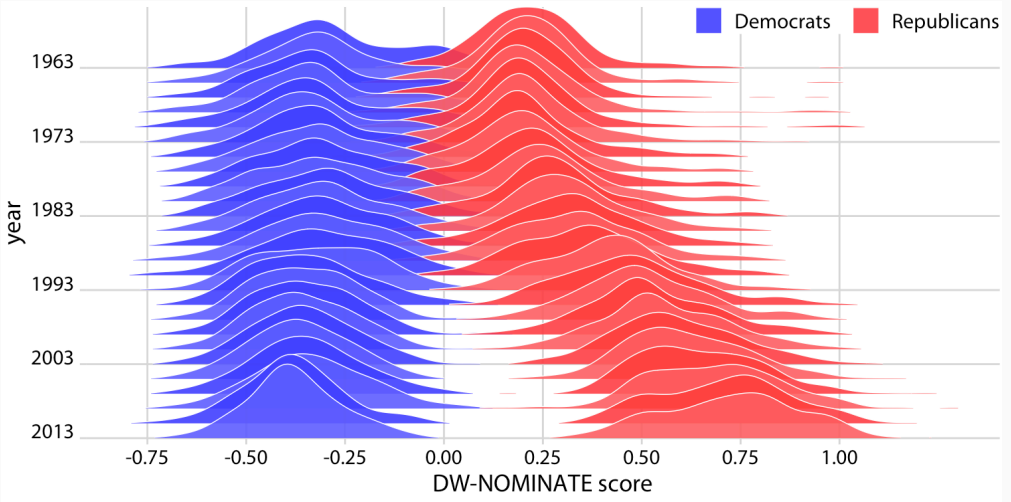
Visualizing many distributions at once (cont.)

- An increasingly popular, compact variant of small multiples for distributions is the **ridgeline plot** (they look like mountain ridgelines).
- The idea is to staggering distributions plots in the vertical direction (i.e., along the horizontal axis).
- Ridgeline plots tend to work particularly well if want to show trends in distributions over time.
- Ridgelines scale to large numbers of distributions.



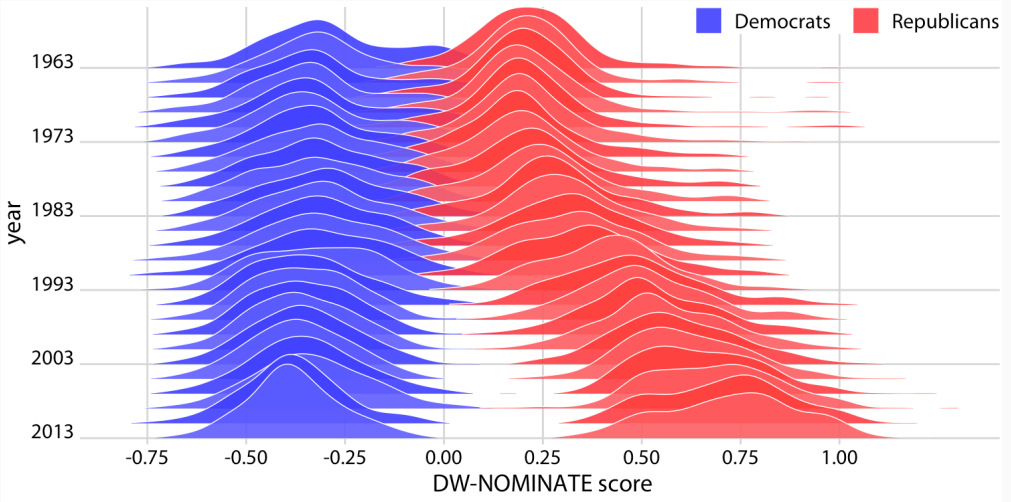
Visualizing many distributions at once (cont.)

- An increasingly popular, compact variant of small multiples for distributions is the **ridgeline plot** (they look like mountain ridgelines).
- The idea is to staggering distributions plots in the vertical direction (i.e., along the horizontal axis).
- Ridgeline plots tend to work particularly well if want to show trends in distributions over time.
- Ridgelines scale to large numbers of distributions.
- Ridgelines can be grouped.



Visualizing many distributions at once (cont.)

- An increasingly popular, compact variant of small multiples for distributions is the **ridgeline plot** (they look like mountain ridgelines).
- The idea is to staggering distributions plots in the vertical direction (i.e., along the horizontal axis).
- Ridgeline plots tend to work particularly well if want to show trends in distributions over time.
- Ridgelines scale to large numbers of distributions.
- Ridgelines can be grouped.
- They are implemented in R with the `ggridges` package.

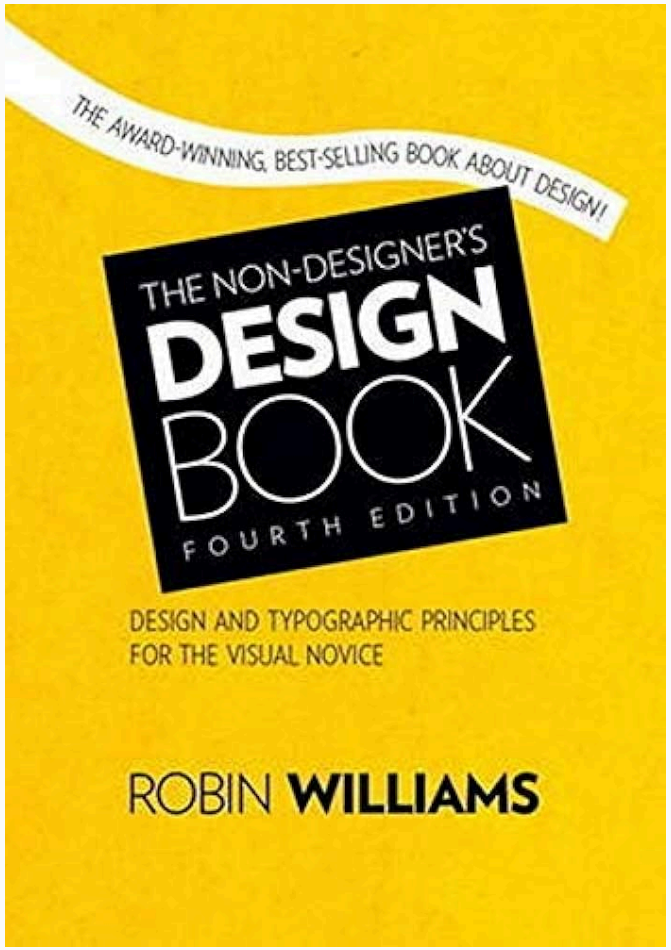


Some principles of graphic design

Graphic design and CRAP

We can this of these **four guiding principles** when creating and critiquing data visuals:

- **Contrast**

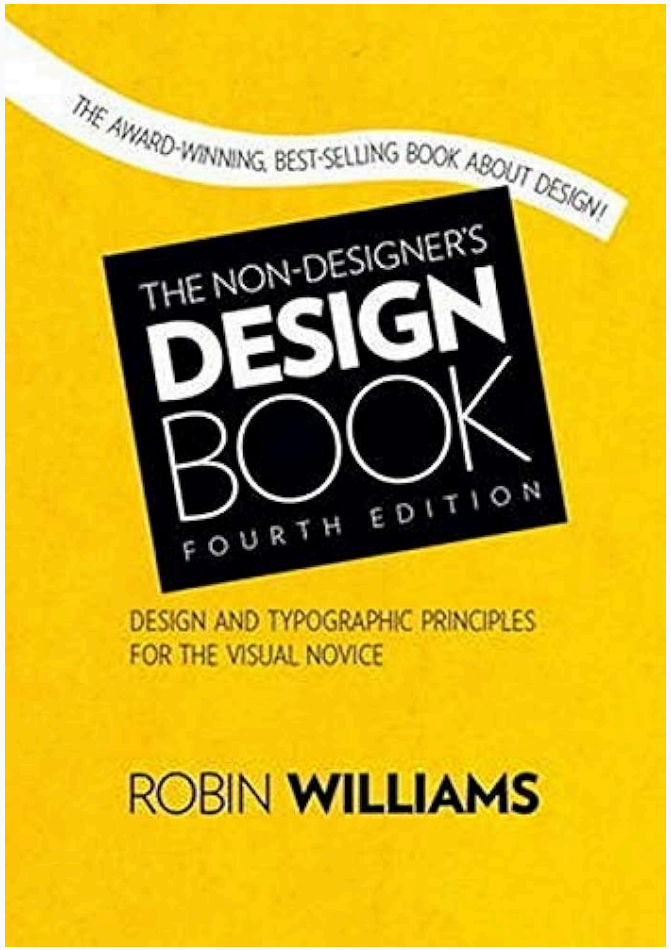


Inspired by Andrew Heiss' Data Visualization with `R` course

Graphic design and CRAP

We can this of these **four guiding principles** when creating and critiquing data visuals:

- **Contrast**
- **Repetition**

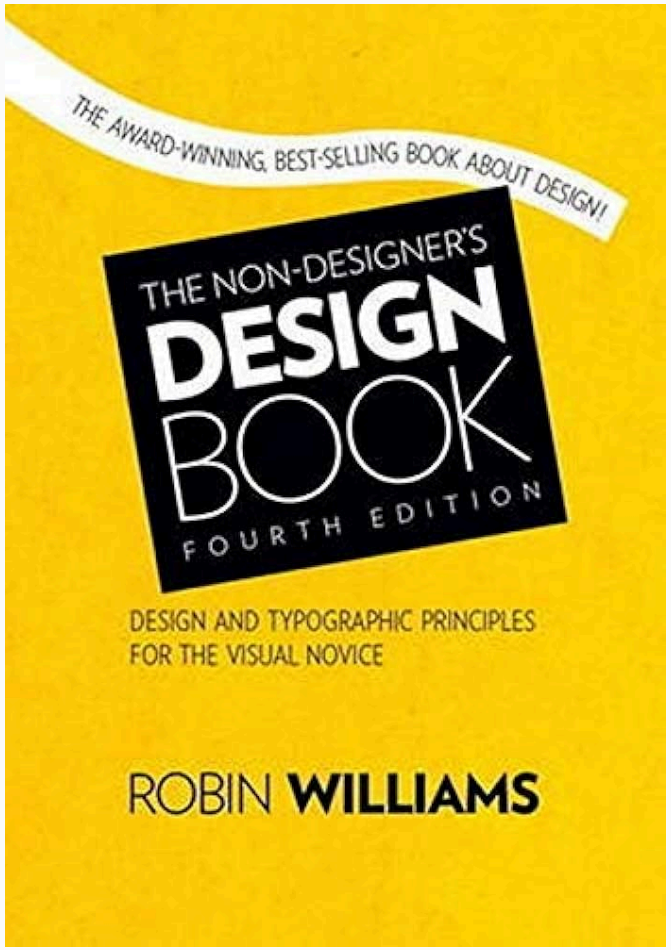


Inspired by Andrew Heiss' "Data Visualization with `R` course

Graphic design and CRAP

We can this of these **four guiding principles** when creating and critiquing data visuals:

- **Contrast**
- **Repetition**
- **Alignment**

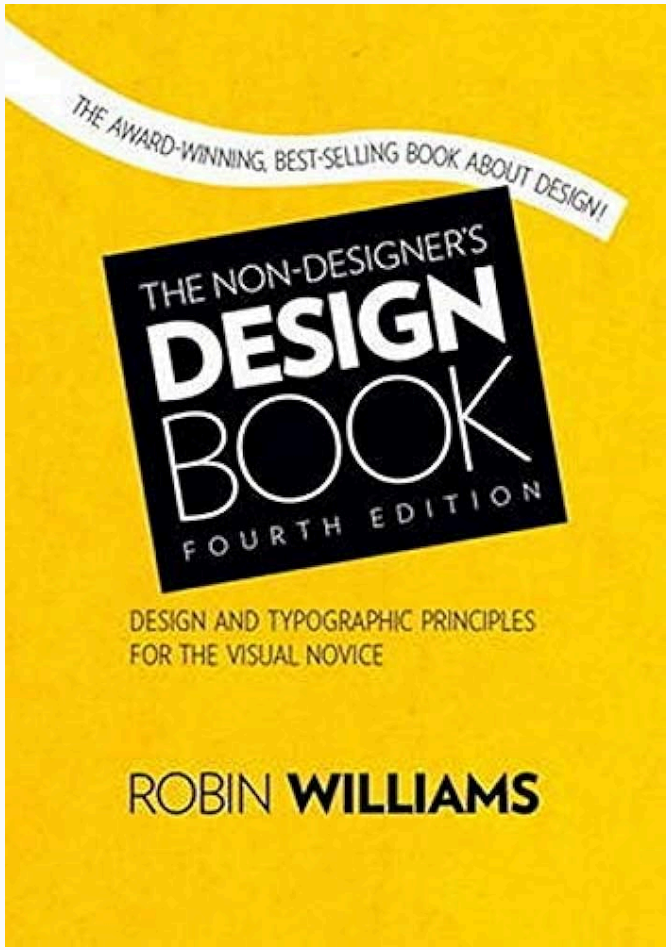


Inspired by Andrew Heiss' "Data Visualization with `R` course

Graphic design and CRAP

We can this of these **four guiding principles** when creating and critiquing data visuals:

- **Contrast**
- **Repetition**
- **Alignment**
- **Proximity**



Inspired by Andrew Heiss' "Data Visualization with `R` course

If two things are not the same, make them different.

Some things to do:

- **Typographic contrast**
  - Type families
  - Weights
  - Size

Serif + Sans serif	*Script* + Serif	**Slab** + Sans serif	etc.
Serif	Sphinx of black quartz, judge my vow		
Sans serif	Sphinx of black quartz, judge my vow		
Slab serif	**Sphinx of black quartz, judge my vow**		
Script	*Sphinx of black quartz, judge my vow*		
Monospaced	Sphinx of black quartz, judge my vow		

If two things are not the same, make them different.

Some things to do:

- **Typographic contrast**
  - Type families
  - Weights
  - Size

Here's a heading

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Here's a heading

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Contrast

If two things are not the same, make them different.

Some things to do:

- **Typographic contrast**
  - Type families
  - Weights
  - Size

Bold + Regular	Regular + Extra light	**Black** + Light	**etc.**
Extra light	Sphinx of black quartz, judge my vow		
Light	Sphinx of black quartz, judge my vow		
Regular	Sphinx of black quartz, judge my vow		
Semi bold	**Sphinx of black quartz, judge my vow**		
Bold	**Sphinx of black quartz, judge my vow**		
Black	**Sphinx of black quartz, judge my vow**		

If two things are not the same, make them different.

Some things to do:

- **Typographic contrast**
  - Type families
  - Weights
  - Size

Here's a heading

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Here's a heading

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Contrast

If two things are not the same, make them different.

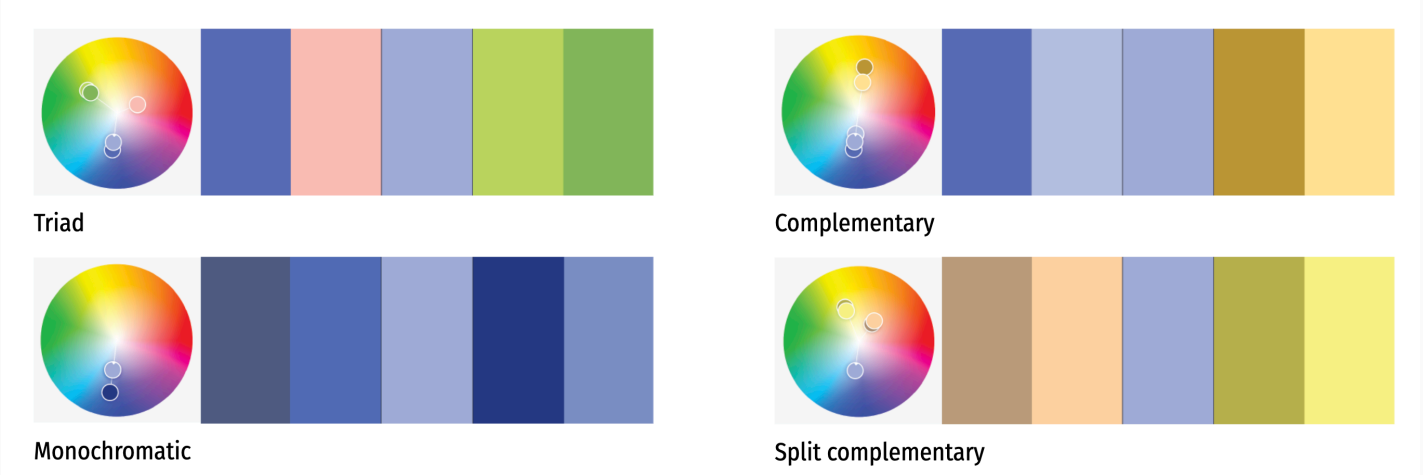
Some things to do:

- **Typographic contrast**
  - Type families
  - Weights
  - Size

If two things are not the same, make them different.

Some things to do:

- **Typographic contrast**
  - Type families
  - Weights
  - Size
- **Color contrast**



We can check Adobe Color

WE HELP PEOPLE BUILD BETTER FUTURES FOR THEMSELVES

Oxfam provides grants and technical support to local organizations around the world to support long-term solutions that help people grow nutritious food, access land and clean water, and—as one of our programs in Jordan illustrates—obtain decent work and fair wages.

MORE THAN A PIPE DREAM

Water scarcity is a major problem in Jordan. Aging water infrastructure and a rapidly increasing population—the conflict in Syria has driven more than 650,000 Syrians to settle in Jordan—have created a situation where every drop counts.

Currently, more than 40 percent of Jordan's water leaks out of broken pipes, so knowing how to fix them is critical. When Oxfam and its partners started a program in northern Jordan to improve the water sector, we made training plumbers—particularly women—a priority.

Funding from the Canadian government helped us equip more than 400 women with basic plumbing skills not only to fix leaks in their homes but to acquire enough plumbing know-how to enter the labor market.

Mariam Tawfeeq Matlaq, 44, picked up a wrench five years ago and started her own business north of Amman soon after. "As soon as I received the training to be a plumber, I had a dream to open a shop," she says, though it wasn't easy to get off the ground. "There are negative perceptions of a woman plumbing in my community. The competition between me and the male plumbers can be difficult."

Still, she says, "I've proved it to people, my community, and the world around me that women can do anything, whether it is conventional or not."

Matlaq has trained many women and recommends them for jobs when she can.

"Women here want to work," she says. "We want opportunities, but often there aren't any for us. We need support from organizations to keep growing these opportunities."

ABOVE, TOP: Mariam Tawfeeq Matlaq fixes the water tank on her rooftop in Zarqa, north of Amman, Jordan, after receiving training from Oxfam on basic plumbing skills. Now she's training other women to become plumbers.

ABOVE, BOTTOM: Matlaq opened her hardware store a year ago in Zarqa and employs several male plumbers who work across the city.

OPPOSITE: "I have been a plumber for five years now. I like it a lot—I especially like the challenges I face," Matlaq says.

PHOTOS: *Abbie Trayler-Smith/Oxfam*

WE SAVE LIVES IN DISASTERS AND CONFLICTS

We work with local organizations to provide assistance during conflicts and disasters, but we also partner with community and national advocates to change the conditions that create them. That's the case in Central America and Mexico, where Oxfam has worked for many decades.

LITTLE CHOICE BUT TO LEAVE

Last fall, Nelson Chavez left his home in El Salvador because he couldn't make enough money to support his family. He worked for a honey producer, bottling honey and selling it from his home. But his income only covered half of what his family needs.

Chavez was one of thousands of people fleeing El Salvador, Guatemala, Honduras, and Mexico to look for a better life in the United States. "What we have in common is the necessity to migrate," he said. "The majority of us do hard work like construction and farm labor, and we are poor. We live on what we make each day."

Oxfam and its partners provided immediate help to Chavez and others in Guatemala with food packages, portable toilets, drinking water, vitamins and rehydration drinks, canopies, and hygiene kits—which included information about how to report acts of violence and human trafficking. In Mexico, we distributed water, thermoses, pots of Vaseline for sore feet, and oral rehydration salts.

But we also provided financial support to local migrants' rights organizations and shelter networks, and called on the governments of Guatemala, Mexico, and the US to protect and guarantee the rights of asylum-seekers and ensure that children aren't separated from their families.

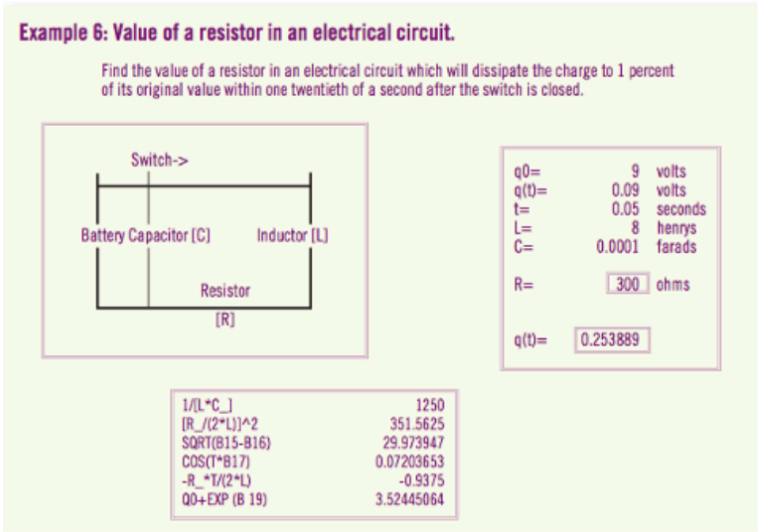
Oxfam President and CEO Abby Maxman visited Tijuana, Mexico, in January to meet with migrants, asylum-seekers, and partner organizations. "The migrants I met in Tijuana are no different than the people who first built our country and what generations of Americans have done: arrive with aspirations to build a better life," she said. "We should live up to our legacy as a welcoming nation that was built on the hard work of immigrants, rather than demonize and criminalize them."

ABOVE, TOP: Nelson Chavez, from El Salvador, left his home and walked to the Guatemala-Mexico border. "There are almost no opportunities to work in my country." *Elizabeth Stevens/Oxfam*

ABOVE, BOTTOM: Oxfam staff Alejandro Orozco and Sherry Toc deliver an inflatable mattress to a man arriving at a shelter in Tecún Umán, Guatemala. *Alyssa Eisenstein/Oxfam*

OPPOSITE: People from La Trinidad, Guatemala, evacuate their community after the eruption of Fuego volcano last June. Oxfam helped those who were displaced. *James Rodriguez/Panos for Oxfam America*

Every item should have a visual connection with something else on the page.



Example 6: Value of a resistor in an electrical circuit.

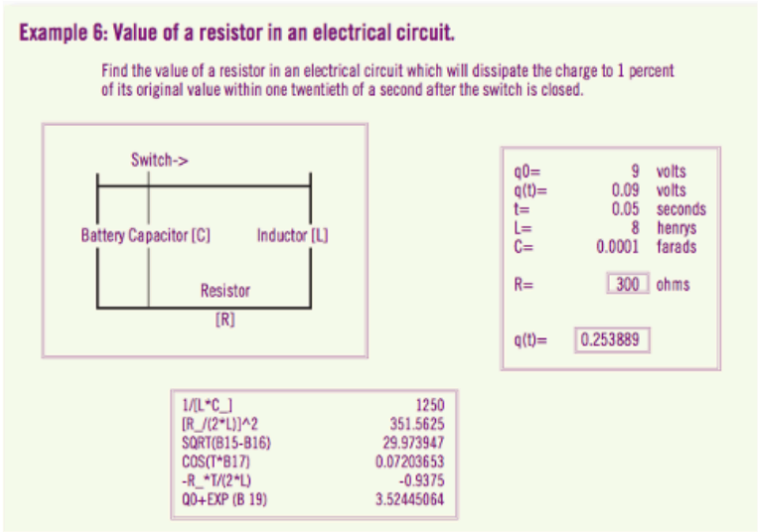
Find the value of a resistor in an electrical circuit which will dissipate the charge to 1 percent of its original value within one twentieth of a second after the switch is closed.

q0=	9	volts
q(t)=	0.09	volts
t=	0.05	seconds
L=	8	henrys
C=	0.0001	farads
R=	300	ohms
q(t)=	0.253889	

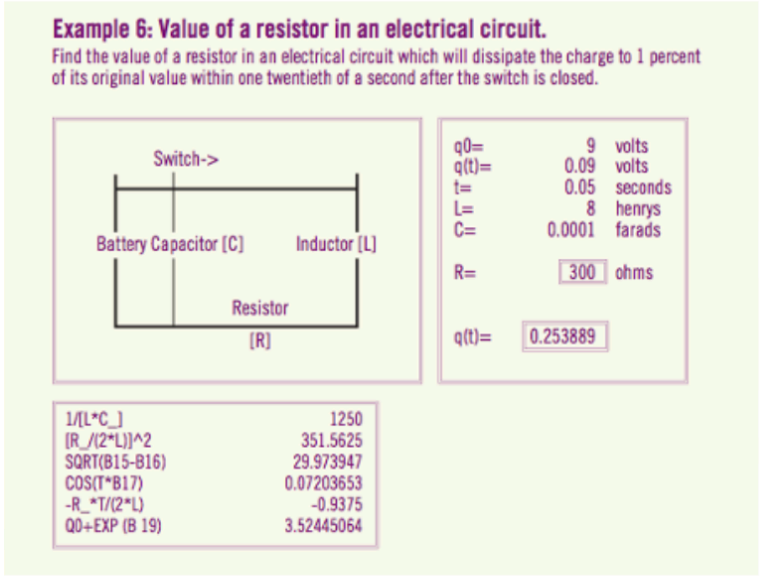
1/[L*C_]	1250
[R_/(2*L)]^2	351.5625
SQRT(B15-B16)	29.973947
COS(T*B17)	0.07203653
-R_*T/(2*L)	-0.9375
Q0+EXP (B 19)	3.52445064

Bad alignment

Every item should have a visual connection with something else on the page.

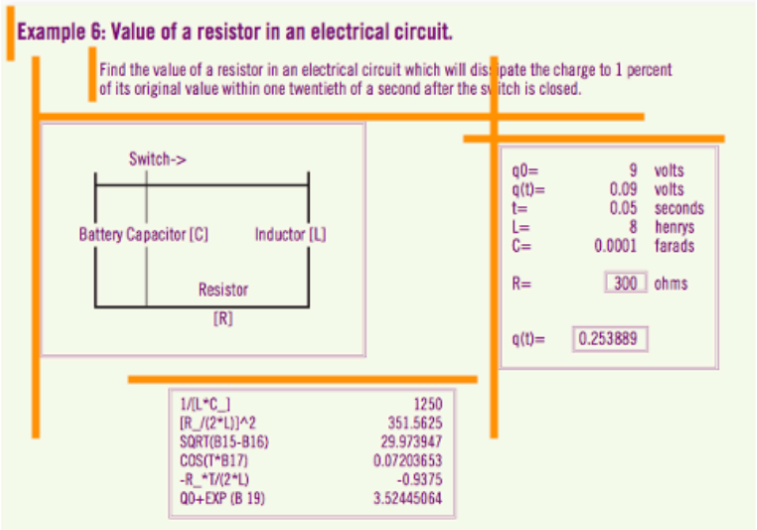


Bad alignment

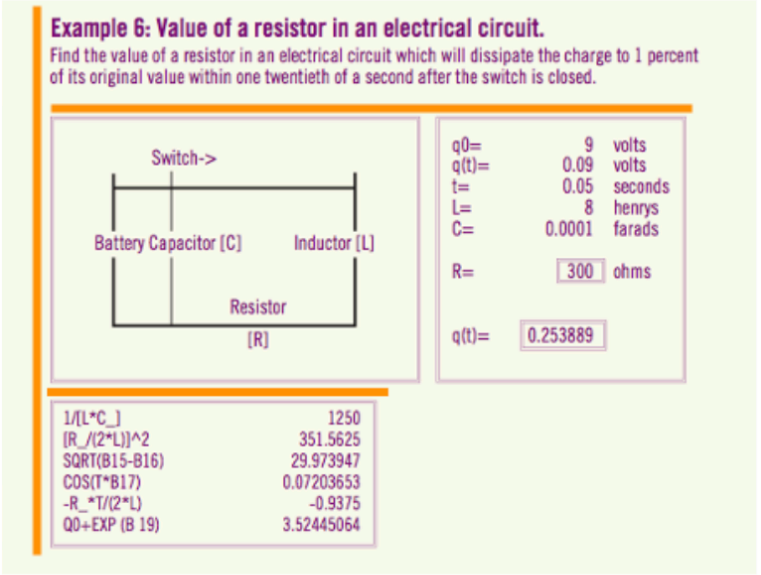


Good alignment—everything is connected to something

Every item should have a visual connection with something else on the page.



4 horizontal alignments; 3 vertical alignments

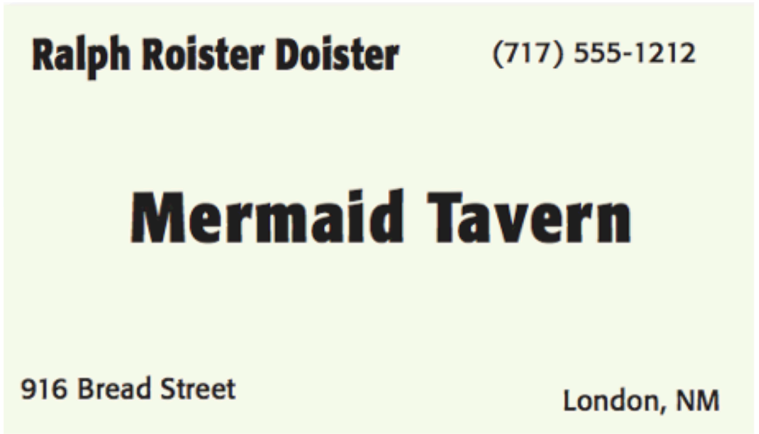


1 shared horizontal alignment; 2 vertical alignments

Proximity

Pull related items together.

Use white space, color, location, contrast, repetition, alignment, etc. to make visually distinct groupings

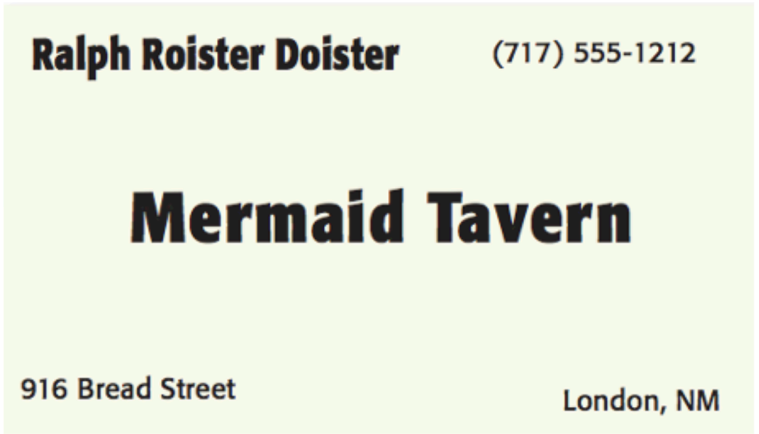


Bad proximity; no logical groupings

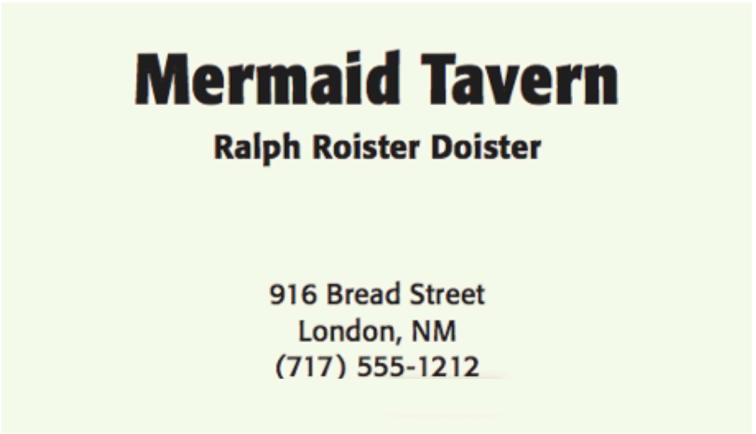
Proximity

Pull related items together.

Use white space, color, location, contrast, repetition, alignment, etc. to make visually distinct groupings

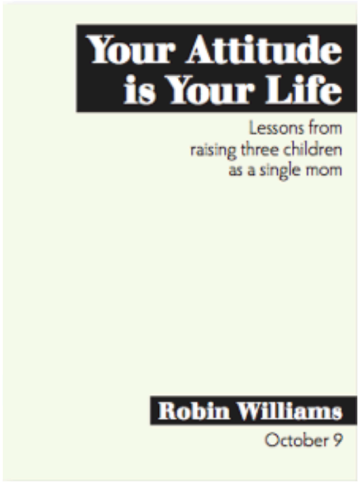


Bad proximity; no logical groupings

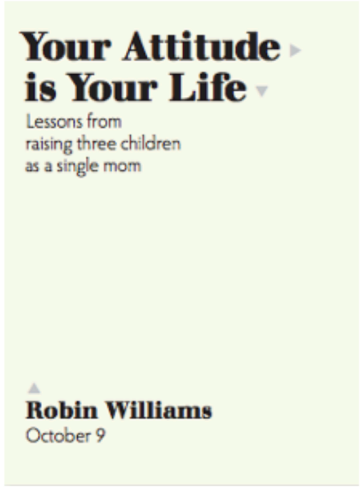


Good proximity; information visually grouped

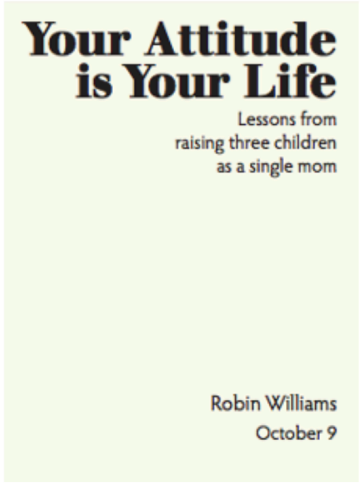
Contrast



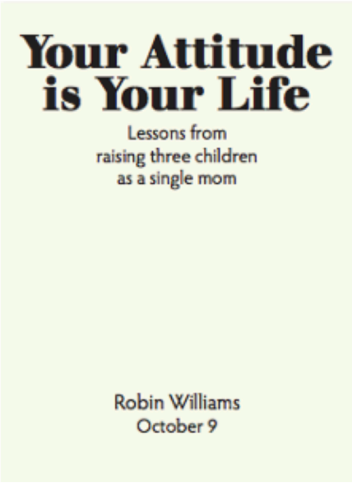
Repetition



Alignment



Proximity



Why data visualization in policy-making?

Data visualization in policy-making

Let's take a couple of minutes to consider some key questions about visualization in the policy context

- In which ways do you think data visualization can be **useful** in the context of policy-making?

Let's take a couple of minutes to consider some key questions about visualization in the policy context

- In which ways do you think data visualization can be useful in the context of policy-making?
- Can you recall an instance where a visual representation of data significantly **influenced one of your (*policy*) decisions**?

Let's take a couple of minutes to consider some key questions about visualization in the policy context

- In which ways do you think data visualization can be useful in the context of policy-making?
- Can you recall an instance where a visual representation of data significantly influenced one of your (*policy*) decisions?
- An argument for data visualizations is that they can help simplify complex data for stakeholders who might not have a technical background. **Do you agree?**

Data visualization in policy-making

Let's take a couple of minutes to consider some key questions about visualization in the policy context

- In which ways do you think data visualization can be useful in the context of policy-making?
- Can you recall an instance where a visual representation of data significantly influenced one of your (*policy*) decisions?
- An argument for data visualizations is that they can help simplify complex data for stakeholders who might not have a technical background. Do you agree? **Are there any risks with simplification?**

Let's take a couple of minutes to consider some key questions about visualization in the policy context

- In which ways do you think data visualization can be useful in the context of policy-making?
- Can you recall an instance where a visual representation of data significantly influenced one of your (*policy*) decisions?
- An argument for data visualizations is that they can help simplify complex data for stakeholders who might not have a technical background. Do you agree? Are there any risks with simplification?
- Do you think data visualization can **improve communication between different departments and with the public**?

Let's take a couple of minutes to consider some key questions about visualization in the policy context

- In which ways do you think data visualization can be useful in the context of policy-making?
- Can you recall an instance where a visual representation of data significantly influenced one of your (*policy*) decisions?
- An argument for data visualizations is that they can help simplify complex data for stakeholders who might not have a technical background. Do you agree? Are there any risks with simplification?
- Do you think data visualization can improve communication between different departments and with the public?
- Are you **currently using data visualization** in your policy work?

Let's take a couple of minutes to consider some key questions about visualization in the policy context

- In which ways do you think data visualization can be useful in the context of policy-making?
- Can you recall an instance where a visual representation of data significantly influenced one of your (*policy*) decisions?
- An argument for data visualizations is that they can help simplify complex data for stakeholders who might not have a technical background. Do you agree? Are there any risks with simplification?
- Do you think data visualization can improve communication between different departments and with the public?
- Are you currently using data visualization in your policy work? **How?**

The case for data visualization in policy

- **Enhanced comprehension and communication**
  - **Simplifier of complex data**: Data visualizations can transform complex datasets into clear, understandable visuals, making it easier for policymakers and stakeholders to grasp intricate information quickly.
  - **Improved communication**: Visuals can bridge communication gaps between technical and non-technical audiences, leveling the accessibility of insights to relevant parties.
- **Facilitator for decision-making**
  - **Trend and pattern identification**: Visual representations can highlight trends, patterns, and anomalies that may not be evident in raw data.
  - **Transparency**: Clear visualizations can help demonstrate the basis for policy decisions.
- **Engagement and participation**
  - **Stakeholder inclusion**: Interactive and visually appealing data presentations can capture the interest of stakeholders, encouraging more active participation in the policy-making process.
  - **Increased understanding**: Well-designed visualizations can educate and inform the public about policy issues, leading to greater awareness and support for policy initiatives.
- **Resource allocation**
  - **Priority setting**: Visual data can help policymakers identify key areas that require attention and allocate resources more effectively and efficiently.
  - **Progress tracking**: Data visualizations can be used to monitor the implementation of policies and measure their changes over time, allowing for adjustments and improvements.

The case for data visualization in policy (cont.)

- **Accountability and transparency**
  - **Reporting**: Designing visualizations make the design of a communication strategy explicit.
  - **Comparison and contrasts**: Visual data can be useful to expose disparities and inequities within groups, prompting action to address these issues.
- **Streamlined processes**
  - **Collaboration**: Visual tools can help different departments and agencies understand each other's data and collaborate more effectively on cross-cutting policy issues.
  - Data harmonization: Visualizations can integrate data from multiple sources, providing a comprehensive view that supports cohesive policy strategies.
- **Computational developments**
  - **Government data**: The increasing availability of large datasets necessitates efficient ways to process and understand data, which visualizations can provide.
    - **Interactivity**: Modern data visualization tools offer interactive features that allow users to explore data dynamically, enhancing their analytical capabilities.
- **Enhanced analytics**
  - **Predictive analysis**: Visualizations can assist in predictive modeling and scenario analysis, helping policymakers anticipate future trends and prepare accordingly.
  - **Management "friendly"**: Visual tools enable rapid analysis and interpretation, which is crucial for timely decision-making in fast-paced policy environments.

Questions?